Mikael Ruohonen

Measurement-Based Automatic Parameterization of a Virtual Acoustic Room Model

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 21.12.2012

Thesis supervisor:

Prof. Ville Pulkki

Thesis advisor:

D.Sc. (Tech.) Juha Merimaa

Aalto University
School of Electrical Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Mikael Ruohonen

Title: Measurement-Based Automatic Parameterization of a Virtual Acoustic Room Model

Date: 21.12.2012 Language: English Number of pages:7+68

Department of Signal Processing and Acoustics

Professorship: Acoustics and Audio Signal Processing Code: S-89

Supervisor: Prof. Ville Pulkki

Advisor: D.Sc. (Tech.) Juha Merimaa

Modern auralization techniques enable making the headphone listening experience similar to the experience of listening with loudspeakers, which is the reproduction method most content is made to be listened with. Room acoustic modeling is an essential part of a plausible auralization system. Specifying the parameters for room modeling requires expertise and time. In this thesis, a system is developed for automatic analysis of the parameters from room acoustic measurements.

The parameterization is based on room impulse responses measured with a microphone array and can be divided into two parts: the analysis of the direct sound and early reflections, and the analysis of the late reverberation. The direct sounds are separated from the impulse responses using various signal processing techniques and used in the matching pursuit algorithm to find the reflections in the impulse responses. The sound sources and their reflection images are localized using time difference of arrival -based localization and frequency-dependent propagation path effects are estimated for use in an image source model.

The late reverberation of the auralization is implemented using a feedback delay network. Its parameterization requires the analysis of the frequency-dependent reverberation time and frequency response of the late reverberation. Normalized echo density is used to determine the beginning of the late reverberation in the measurements and to set the starting point of the modeled late field. The reverberation times are analyzed using the energy decay relief.

A formal listening test shows that the automatic parameterization system outperforms parameters set manually based on approximate geometrical data. Problems remain especially in the precision of the late reverberation equalization but the system works well considering the relative simplicity of the processing methods used.

Keywords: room acoustics, auralization, virtual acoustics, room impulse response, acoustic source localization, reverberation

AALTO-YLIOPISTO
SÄHKOTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Mikael Ruohonen

Työn nimi: Mittauksiin perustuva huoneakustisen mallin automaattinen parametrisointi

Päivämäärä: 21.12.2012 Kieli: Englanti Sivumäärä:7+68

Signaalinkäsittelyn ja akustiikan laitos

Professuuri: Akustiikka ja äänenkäsittely Koodi: S-89

Valvoja: Prof. Ville Pulkki

Ohjaaja: TkT Juha Merimaa

Modernien auralisaatiotekniikoiden ansiosta kuulokkeilla voidaan tuottaa kuuntelukokemus, joka muistuttaa useimpien äänitteiden tuotannossa oletettua kaiutinkuuntelua. Huoneakustinen mallinnus on tärkeä osa toimivaa auralisaatiojärjestelmää. Huonemallinnuksen parametrien määrittäminen vaatii kuitenkin ammattitaitoa ja aikaa. Tässä työssä kehitetään järjestelmä parametrien automaattiseksi määrittämiseksi huoneakustisten mittausten perusteella.

Parametrisaatio perustuu mikrofoniryhmällä mitattuihin huoneen impulssivasteisiin ja voidaan jakaa kahteen osaan: suoran äänen ja aikaisten heijastusten analyysiin sekä jälkikaiunnan analyysiin. Suorat äänet erotellaan impulssivasteista erilaisia signaalinkäsittelytekniikoita käyttäen ja niitä hyödynnetään heijastuksia etsivässä algoritmissa. Äänilähteet ja heijastuksia vastaavat kuvalähteet paikannetaan saapumisaikaeroon perustuvalla paikannusmenetelmällä ja taajuusriippuvat etenemistien vaikutukset arvioidaan kuvalähdemallissa käyttöä varten.

Auralisaation jälkikaiunta on toteutettu takaisinkytkevällä viiveverkostomallilla. Sen parametrisointi vaatii taajuusriippuvan jälkikaiunta-ajan ja jälkikaiunnan taajuusvasteen määrittämistä. Normalisoitua kaikutiheyttä käytetään jälkikaiunnan alkamisajan löytämiseen mittauksista ja simuloidun jälkikaiunnan alkamisajan asettamiseen. Jälkikaiunta-aikojen määrittämisessä hyödynnetään energy decay relief -metodia.

Kuuntelukokeiden perusteella automaattinen parametrisaatiojärjestelmä tuottaa parempia tuloksia kuin parametrien asettaminen manuaalisesti huoneen summittaisten geometriatietojen pohjalta. Järjestelmässä on ongelmia erityisesti jälkikaiunnan ekvalisoinnissa, mutta käyttetyihin suhteellisen yksinkertaisiin tekniikoihin nähden järjestelmä toimii hyvin.

Avainsanat: huoneakustiikka, auralisaatio, virtuaaliakustiikka, huoneen impulssivaste, akustinen paikannus, jälkikaiunta

Preface

First of all, I would like to thank Veronique Larcher and the rest of the Sennheiser Research and Innovation for the opportunity to work in an innovative and highly talented team. Thanks to all my co-workers for a wonderful experience and for the new insights to audio technology I got every day. I owe a lot especially to those who worked on related projects and gave me important information on the context of my work.

My biggest thanks go to my instructor Juha Merimaa who worked with me on the project, gave feedback on my thesis and helped me with his expertise. Thank you also to my supervisor Ville Pulkki for creating bridges between Espoo and California and for supervising my work.

Lastly, I would like to thank my family and friends for all the oversea support during my stay in California and back in Finland.

Espoo, December 21, 2012

Mikael J. Ruohonen

Contents

Abbreviations

BRIR	binaural room impulse response
GPS	global positioning system
RIR	room impulse response
ISM	image source model
FEM	finite-element method
BEM	boundary-element method
FDTD	finite difference time domain
ART	acoustic radiance transfer
BRDF	bidirectional reflectance distribution function
FDN	feedback delay network
HRTF	head-related transfer function
MLS	maximum length sequence
EDC	energy decay curve
EDT	early decay time
EDR	energy decay relief
STFT	short-time Fourier transform
NED	normalized echo density
SRP	steered response power
IMP	incremental multi-parameter
MVDR	minimum variance distortionless response
DOA	direction of arrival
MUSIC	multiple signal classification
TDOA	time-difference of arrival
GCC	generalized cross-correlation
AMDF	average magnitude difference function
CC	cross-correlation
PHAT	phase transform
SCOT	smoothed coherence transform
ML	maximum likelihood
AMSF	average magnitude sum function
MAMDF	modified average magnitude difference function
TDE	time difference estimation
LS	least squares
SRP-PHAT	steered response power using phase transform
SIRR	spatial impulse response rendering
RMS	root mean square
MP	matching pursuit
MUSHRA	multi-stimuli hidden reference and hidden anchor
GUI	graphical user interface

1 Introduction

Headphones are widely used for reproduction of all types of audio material. Most content is, however, mixed and mastered in a control room with speakers. In headphone listening, sounds are usually localized inside the listener's head which makes the listening experience very different from that achieved with loudspeakers. Surround sound content causes even bigger problems in headphone reproduction as the spatial impression of several speakers is lost when the channels are downmixed to stereo. In order to achieve a listening experience closer to that experienced with a speaker setup in a real room, this listening environment can be simulated artificially.

The process of creating an artificial spatial audio impression is called auralization [1]. There are various possible approaches for auralization. A relatively simple way is to make a binaural recording at the ears of a listener and play it back through headphones. Often the binaural room impulse responses (BRIR) are recorded to enable the use of any source material by convolution of the source signal and the measured BRIRs. This approach is, however, tied to specific source and listener positions and to a specific head. Modern computational modeling approaches enable the separation of the listener and his or her position. In addition, auralization is not limited to real-world, measured set-ups because it is possible to create arbitrary, virtual acoustic environments. These flexible, dynamic systems also allow interaction with the user. Especially tracking the listener's movement is important for the spatial impression as the sound sources keep their positions even if the listener moves his or her head. Moving of the head also helps in the localization of the sources.

The virtual acoustic modeling process can be divided into three key parts: the source, the medium and the receiver [2]. The source radiates sound into the environment following its characteristic radiation pattern that depends on frequency of the transmitted signal. Sound propagates from the source to the listener through various paths in the surrounding space depending on the positions of the walls and other obstacles as well as on their materials and other properties. Finally, a human listener hears a direction-of-arrival-dependent modified version of the sound.

This thesis focuses on the middle part of the modeling problem. An existing room acoustic model is extended with an automatic parameterization system. Without such automatic parameterization, configuring the room model requires a lot of manual work and expertise in room acoustics. The goal of this work is to develop a system for measuring a room and automatically determining the parameters of the model. In addition to reducing manual work, automatic parameterization might enable more accurate reproduction of existing rooms.

In order to achieve a plausible result, room acoustic measurements and various signal processing procedures are required in the parameterization system. The locations of the sound sources and reflection surfaces must be analyzed. Several techniques for solving similar problems have been developed in the general context of source localization which is important in topics ranging from global positioning system (GPS) to telecommunications, modern conference systems, underwater acoustics and radar. The analysis of the reverberant field in a room is strongly

related to the long history of architectural acoustics. The purpose of the entire analysis system is to bring more realism into the increasingly wide set of applications of auralization and virtual acoustics.

1.1 The Existing Auralization System

The structure of the auralization system for which the automatic parameterization system is developed is depicted in Figure 1. The main functionalities of the auralization system are implemented on an embedded digital signal processing platform. The system takes in multichannel audio and gives a binaural signal for headphone reproduction as output. The inputs are fed to separate processing blocks for the direct sound, early room reflections and late reverberation.

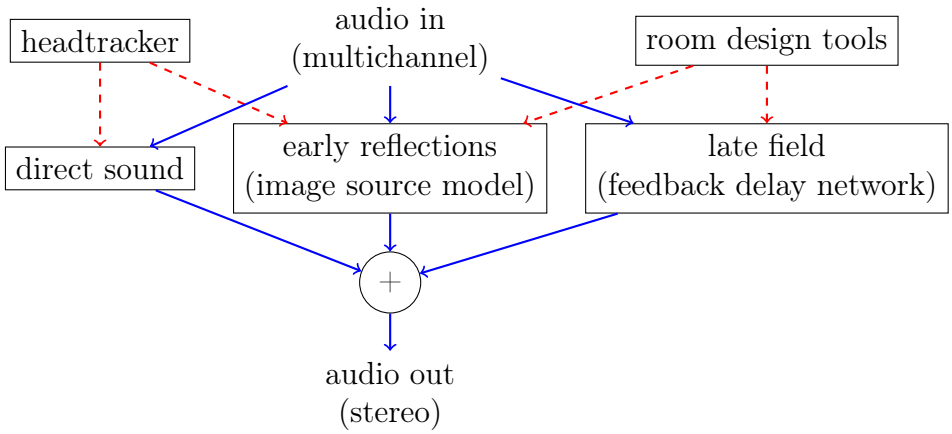

Figure 1: Functional structure of the auralization system. Solid blue and dashed red arrows denote audio and data streams, respectively.

The direct sound and early reflections are auralized using head-related transfer functions (HRTFs) to create a sensation that they arrive from the desired directions. An HRTF represent the ratio of the frequency-dependent sound pressure at the listener's ear and at a specific point in the free field [3]. HRTFs can be measured at the listener's ears for different arrival angles. These transfer functions can be presented with filters that can be applied to sound signals in order to achieve the sensation of the sound arriving from the direction of the measured HRTF [4]. The measurement of individual HRTFs requires time and special equipment. A general set of HRTFs can be used for different listeners in auralization at the risk of giving suboptimal performance for some listeners.

The auralization system responds to head rotations by using head orientation data from a headtracker in the headphones. The orientation of the head is compensated for in the HRTF processing of the direct sounds and early reflections, keeping the synthesized sound sources and the room still when the head is rotated. Therefore, it is possible for the user to actually have a sensation of being in the virtual

space and at the same time the localization of virtual sources becomes more accurate since head movements provide additional localization cues.

In this work, the existing auralization system is used as is and only the block "room design tools" in Figure 1 is modified to include an automatic parameterization based on measurements. The room design is done using a set of Matlab tools that provide parameters for each processing block for run-time processing of the audio input. The addition of the Matlab-based automatic parameterization using measurement data is made within the limitations of the existing room model and, in addition, the HRTF processing, headtracking and other parts of the auralization system were not to be modified in this work.

Although an auralization system like the one described here has several applications, this system was originally created for headphone listening of multichannel audio simulating a loudspeaker setup in a room. It is also the application for which the automatic parameterization system was designed for and thus explains several decisions made throughout this work, especially the emphasis on the perceptual performance of the auralization system instead of on the theoretical precision of individual parts of the analysis system. It is also important to note that the aim of the new system was to work for static listener and loudspeaker positions in the room, a constraint which steered the development of the analysis from the beginning.

1.2 Thesis Structure

The thesis is organized as follows. Basic theory of room acoustics is reviewed in Chapter 2. Different room acoustic modeling techniques are presented in Chapter 3. Typical room acoustic measurements and calculations based on them are presented in Chapter 4. An overview of the field of acoustic source localization is given in Chapter 5 with an emphasis on methods related to this work. The following two chapters describe the room acoustic analysis and room model parameterization. The analysis and parameterization of the direct sound and early reflections are presented in Chapter 6 whereas the processing related to the late reverberation is described in Chapter 7. A listening test was conducted for the evaluation of the implemented automatic parameterization system. The test and its results are described in Chapter 8 along with discussion on the test and the system's performance. The current work is concluded in Chapter 9.

2 Basics of Room Acoustics

2.1 Sound Propagation in Enclosed Spaces

The listening environment is an important part of the listening experience. Anyone can tell apart a large concert hall from an acoustically treated listening room or a small but reverberant bathroom. Moreover, different rooms are suitable for different listening. Because of its importance, room acoustics has been widely studied. Much of the knowledge on room acoustics is based on know-how of concert hall and studio design. A room is an acoustically complex system which is often studied and described with statistical measures although exact geometrical and physical information can nowadays be more extensively used in room analysis, design and auralizations because of the high computational power and highly developed modeling tools available.

In an acoustical free field, sound propagates as longitudinal pressure waves from a sound source according to the radiation pattern of the source. The propagation of sound in air and other lossless fluids is described by the wave equation which, using one spatial dimension, is [5]:

$$c^2 \frac{\partial^2 p}{\partial x^2} = \frac{\partial^2 p}{\partial t^2}, \tag{1}$$

where c is the speed of sound, which depends on the medium and the temperature, p is the sound pressure, x is the location in one dimension and t is the time.

The basic wave types which are often assumed in most analysis of acoustic fields are spherical waves and plane waves. These refer to the shape of the wavefront, i.e. the points of the propagating sound having the same phase. An ideal point source emits spherical waves to its surroundings. The pressure of the spherical wave is [5]

$$p = \frac{\mathrm{i}\omega\rho_0}{4\pi r} Q \exp^{[\mathrm{i}(\omega t - kr)]}, \tag{2}$$

where i is the imaginary unit, ω is the angular frequency, ρ_0 is the mean density of air, r is the distance from the source, Q is the amplitude of the volume velocity function and $k = \omega/c$ is the wave number. Here it can be seen that the pressure of the spherical wave is proportional to the inverse of the distance. An ideal point source is infinitely small but a source can be approximated as a point source if it is relatively small compared to the wavelength. In the far field of the source, a small segment of the spherical wave can be approximated with a theoretical plane wave.

The sound arriving to the receiver, or the listener, is completely described by the source and the medium the sound propagates through. In practice, apart from anechoic chambers which even themselves are approximations of a free field, we are surrounded by obstacles that obstruct the propagation of the sound waves and modify propagation paths. Reflections from these obstacles create secondary propagation paths between the source and the receiver. Outdoors there is possibly only one surface, the ground, causing a single reflection. In a room, the situation is much more complex. Sound reflects from all the boundaries of the room creating several new propagation paths between the source and the receiver. The sound waves radiating from the source to different directions also reflect from one wall or obstacle

to another until the losses in the air and the materials make the vibrations small enough to become inaudible. The wavefronts traveling back and forth in the room reach the listener at different times and energies creating the acoustical experience characteristic to that room.

2.2 The Room Impulse Response

A room impulse response (RIR) represents the pressure changes at a receiver point as a function of time after an impulse has been emitted at a transmitter point in the room. In theory, it characterizes entirely the propagation path and room effects between the two points. The RIR can be divided into three parts [2] which can be seen in Figure 2. First, there is the direct sound propagating through air from the transmitter to the receiver if there are no obstacles between the two. In the RIR, the direct sound can be seen as a strong peak at the beginning. The direct sound arrives at a time corresponding to the distance between the two points. After the direct sound, the RIR has sparsely spaced lower peaks. These peaks are due to specular reflections from walls and other obstacles. Their arrival times depend on the room geometry and positions of the transmitter and the receiver in the room. Sound waves keep propagating in the room reflecting from one wall to the other, spreading the sound somewhat evenly to the whole room. This can be seen as an increased density of arriving impulses in the RIR. The amplitudes of single impulses get smaller with time since sound has to travel longer and is attenuated due to losses in air and wall materials. As the density of the arriving reflections increases, the RIR reaches the late reverberation tail which is close to exponentially decaying noise. This is the third part, called the late field. It can often be treated statistically as is described later.

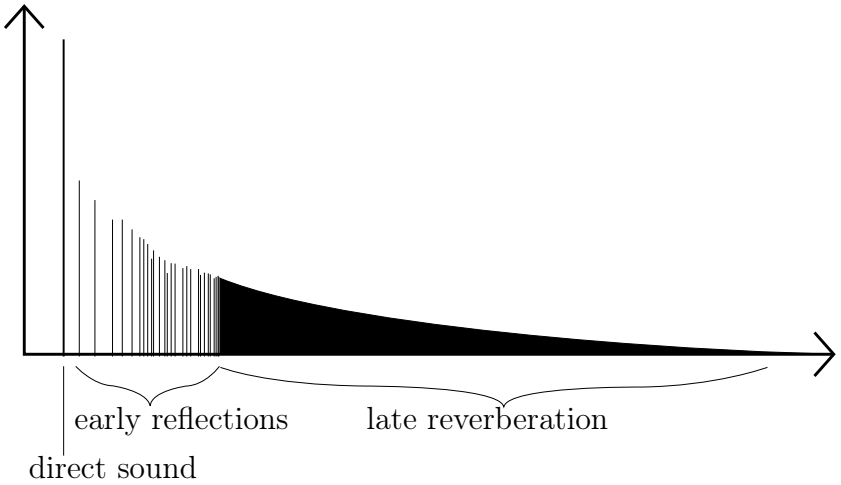

Figure 2: Structure of a typical room impulse response.

2.3 Effects of Room Boundaries

Part of the energy of the sound wave is transformed into heat as it propagates through air and different materials. This phenomenon is called absorption. Part of the energy reaching the walls also passes to the other side. From the point of view of the room this can also be considered as absorption although the sound is not absorbed and continues to contribute to the sound field of the space on the other side of the wall. Absorption is frequency-dependent which means that air and material absorption not only decrease the total level of the sound but also have an effect on the relative frequency content. The frequency-dependent absorption is different for different materials which makes the materials in the walls an important factor in the way the room sounds. Hard and heavy concrete walls reflect almost all of the energy of the sound wave and thus have very little absorption at all frequencies. Lighter, smooth walls reflect higher frequencies without much losses but lower frequencies make them vibrate, absorbing some of the energy at these frequencies. Porous materials cause losses at all frequencies, especially in the high frequency range. Reflection absorption is a complex phenomenon depending on the properties of the arriving wave and the structure and impedance of the material. Assuming a diffuse field where all arrival directions are equally likely and sound waves reach the enclosure boundaries at random phases, absorption at a boundary can be simplified to a simple frequency-dependent absorption coefficient $\alpha(f)$ which is the fraction of the energy lost in the reflection [5]. Absorption can be seen in the RIR as attenuated reflection peaks and as spreading of the reflected impulses.

Wall materials are rarely totally smooth and in addition to small- or large-scale ripple or roughness of the surface, there are often smaller obstacles, such as lights, bookshelves and paintings on the wall making the total profile of the wall uneven. All these irregularities with finite extension cause the arriving sound waves to spread in various directions instead of one specular reflection. A reflection where the energy is spread to different directions is called a diffuse reflection. This scattering of sound is caused by diffraction which refers to edges and irregular shapes functioning as secondary sound sources. Like absorption, scattering is frequency-dependent. For wavelengths that are small compared to the dimensions of the irregularities, scattering is stronger whereas sound waves with wavelengths large compared to the dimensions may reflect specularly even from rough surfaces. Irregular surfaces have frequency- and angle-dependent radiation patterns. Scattering may also occur with curved smooth surfaces [5]. Special shapes are used to create maximally diffusive reflections. Because of diffuse reflections and diffraction, the early part of the RIR does not consist only of separate early reflections but also of small ripple between these distinct peaks.

2.4 Wave Field Theory

There are two different approaches used for mathematical inspection of room acoustics which will be important in the following chapter where room acoustic modeling methods are described. The first one is wave field theory and it is based on the fun-

damental laws of sound described by the wave equation. The importance of wave field theory in general room acoustic observation lies in the theory of eigenmodes of the room. The wave equation can be solved in closed form for simple geometries and numerically for more complex cases. The solutions to the equations are called eigenfunctions [5]. These are sinusoidal functions representing characteristic standing waves, or normal modes, of the room at frequencies called eigenfrequencies. The average density of eigenfrequencies is [5]

$$\frac{\mathrm{d}N_f}{\mathrm{d}f} = 4\pi V \frac{f^2}{c^3}, \tag{3}$$

where $\frac{\mathrm{d}N_f}{\mathrm{d}f}$ is the number of eigenmodes per Hz and V is the room volume. The mode density is proportional to the room volume and to the square of the frequency. Hence, the separation of eigenfrequencies is larger in smaller rooms and at lower frequencies. The individual modes become insignificant when the average spacing of the eigenfrequencies is less than one third of the bandwidth of an eigenmode [5]. This frequency limit is denoted by the Schroeder frequency [6]

$$f_{\mathrm{Schroeder}} = 2000\sqrt{\frac{RT_{60}}{V}}, \tag{4}$$

where RT_{60} is the room reverberation time, i.e. the time it takes for sound to decay 60 dB. The room volume is in the denominator, which means that in large halls individual eigenmodes do not have a strong effect whereas in small rooms they need to be individually studied.

2.5 Geometrical Acoustics

The second approach to understanding sound fields in rooms is based on geometrical observation of the propagation of sound. In geometrical acoustics sound is assumed to behave like rays of light. This can be done when the room dimensions are significantly larger than the observed wavelengths [5]. The key concepts in geometrical acoustics are the laws of reflection similar to optics and the decay of the wave with increasing distance. The reflections of sound rays at rigid boundaries obey the rule that the angle of incidence equals the reflection angle. The absorption in the boundary material can be taken into account as a loss of energy proportional to the absorption coefficient. The average density of reflections arriving to a receiver point at the time t is [5]

$$\frac{\mathrm{d}N_t}{\mathrm{d}t} = 4\pi \frac{c^3 t^2}{V}. \tag{5}$$

It can be seen that the echo density increases proportional to the square of time.

2.6 The Mixing Time

The point in time where the individual echoes do not matter anymore and the diffuse late field has been reached is called the mixing time. It can be seen as a time

domain counterpart of the Schroder frequency. Jot et al. [7] derive a formula for the mixing time based on Polack's [8] studies where ten reflections within a typical time resolution of the auditory system, here 24 ms, is assumed to be enough to create a perceptual late field:

$$t_{\text{mixing}} = \sqrt{V} \quad [\text{ms}]. \tag{6}$$

Rubak and Johansen [9] present another approach to the mixing time calculation using a common room acoustic measure called the mean free path. It is the average path length sound travels between two reflections. The mean free path can be calculated as a function of room volume V and total surface area S:

$$l_{\text{m}} = 4\frac{V}{S}. \tag{7}$$

The diffuse late field is assumed in the calculation of the mean free path. In other words, all directions of the sound propagation must be equally likely and sound energy has to be distributed equally around the room for average path length to be a sensible measure [5]. The mean free path -based mixing time calculation, however, relies on the assumption that the late field will be approximately diffuse when sound waves have reflected a large enough number of times on average [10]. Assuming four reflections is sufficient, the mixing time becomes [10, 9]

$$t_{\text{mixing}} = 4l_{\text{m}}\frac{10^3}{c} = 47\frac{V}{S} \quad [\text{ms}] \tag{8}$$

Lindau et al. [10] compare the different mixing time predictors with perceptual tests. Based on the listening tests, they derive regression formulas for using the square of the volume of the room (see Equation (6)), the mean free path (see Equation (8)), plain volume of the room and reverberation time (as suggested by Hidaka et al. [11]) in the prediction of the mixing time. According to the listening test, the mean free path -based prediction works best and the regression formula derived for the mean free path is

$$t_{\text{mixing}} = 20\frac{V}{S} + 12. \tag{9}$$

2.7 The Late Field

The late field is the part of the reverberation where individual reflections become insignificant and the diffuse sound field is dominant. As mentioned earlier, in a diffuse field sound energy is assumed to be equally distributed throughout the volume of the room and all directions of wave propagation are assumed equally likely. In other words, in the statistical approach to the late field of the room, the mixing time is assumed to be reached which makes it possible to neglect the discrete reflections and propagation directions. The late field assumptions are often made, even though they are approximations, especially with smaller rooms, because they enable the room to be studied statistically.

The basis for the theory of room reverberation was laid by Sabine's studies [12]. His empirically deduced formula on the connection between reverberation time RT_{60},

i.e. the time it takes for interrupted wideband sound to decay 60 dB, room volume V and equivalent absorption area A has been widely used ever since and can also be deduced analytically [5]. The Sabine formula is (in room temperature):

$$RT_{60} = 0.161\frac{V}{A}, \tag{10}$$

where A is defined as:

$$A = \sum_i S_i \alpha_i, \tag{11}$$

where S_i and α_i are the area and absorption coefficient of the surface i, respectively. If frequency-dependent values of the absorption coefficients are used, the reverberation time can also be calculated in frequency bands.

3 Room Acoustic Modeling

The problem of room acoustic modeling for auralization and virtual acoustic environment purposes has been approached from many directions. A basic distinction can be made between *physical* and *perceptual* room acoustic modeling [13]. In physical modeling, the acoustics are modeled based on the geometrical knowledge of the room and the positions of the source and the listener. Sound propagation in the room is modeled at a low level considering individual wave fronts and their reflections. Perceptual modeling aims at a perceived impression similar to the room being modeled but the implementation is often an efficient algorithm rather than a precise model. The room model used in this work follows a common approach to model the early part of the RIR using a physical method and the late reverberation tail using a perceptually-motivated artificial reverberation algorithm. In this chapter, physical room acoustic modeling methods and artificial reverberation methods are presented and the room model using both approaches is introduced.

3.1 Physical Room Acoustic Modeling

Physical room acoustic modeling techniques include scale modeling and computational simulation [1]. All the techniques described in this section belong to the latter group. Scale models require building the actual models and using scaled-down sources, receivers and sound waves whereas computational models can be easily modified and reused in another computational device. Thus, computational models are considerably more practical for auralization purposes. An overview of computational room acoustic modeling methods is presented in Figure 3. They can be divided into three categories [2]: wave-based methods, ray-based methods and statistical methods. Statistical methods are not suitable for auralization [2] and thus are not described in this thesis. Wave-based and and ray-based methods are explained below with an emphasis on the image source model (ISM) which is used in an existing implementation that the current work attempts to parameterize.

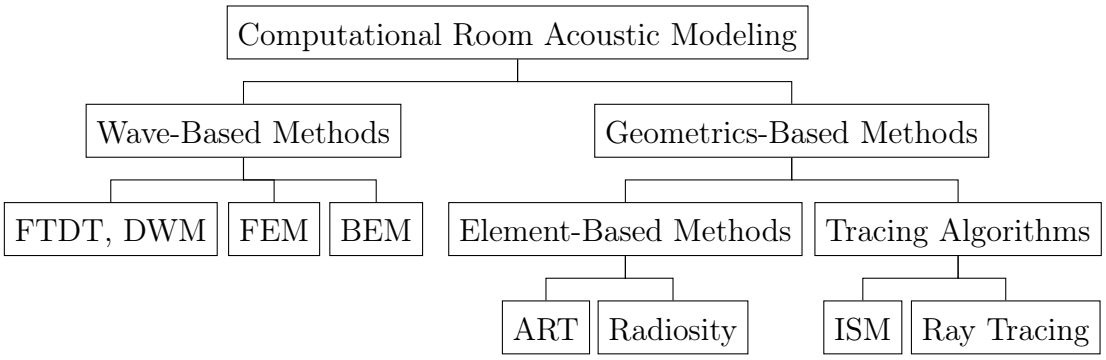

Figure 3: Physics-based room acoustics modeling methods. Adapted from [14].

3.1.1 Wave-Based Methods

Accurate room acoustic modeling requires solving the wave equation (1). Solving the equation in the analytical form is possible only in the case of simplest room geometries and in practice numerical methods are used [15]. Typically the room volume or boundaries are divided into a finite number of elements for which the computations are performed. Room size and the highest modeled frequency are restricted in the wave-based methods because there needs to be 6-10 elements per shortest wavelength and hence the matrices grow quickly with increasing frequency or room size [16].

The finite-element method (FEM) and the boundary-element method (BEM) are based on solving the wave equation in finite elements across the volume or the boundaries of the room, respectively [16]. In FEM, the volume of the room is divided into elements and the wave equation is solved individually in these so called nodes. Using the Kirchhoff-Helmholtz integral equation, the pressure in a cavity can be solved if both pressure and particle velocity are known at the boundaries [16]. This principle is used in BEM where the boundaries of the space are divided into elements.

Finite-difference time domain (FDTD) methods are another approach to solving the wave equation numerically, but in their case in the time domain [17]. The gradients or derivatives of the wave equation are modeled with finite differences and evaluated at each element in the volume. Different boundary conditions can be set to take into account the effects of the wall impedances [18]. Similar wave equation approximation can be done with a mesh of digital waveguides (DWM, digital waveguide mesh) [19]. The simple waveguide mesh approach has several problems with arbitrary room shapes and dispersion [14]. Several improvements are available using, for instance, interpolation techniques [20].

3.1.2 Ray-Based Methods

As was discussed earlier, the wave phenomena are not of high importance above the Schroeder frequency and thus the wave-based methods are often not worth the computational load at higher frequencies and more efficient geometrical approaches can be used. Ray-based or geometrical-based methods assume sound to behave like rays of light [21]. Some of the effects excluded by the principles of geometrical acoustics, such as diffraction, can be adopted to these techniques [22, 23] but they are still rough approximations at low frequencies.

Ray-Tracing

Ray-tracing is a widely used method in architectural acoustics software for predicting the behavior of sound in rooms [24]. It is based on sending rays of sound to random directions from the source and tracing their paths through the space modeling their reflections. The receiver is modeled as a volume and all rays passing through the volume are added to received energy [14]. Ray tracing can be implemented efficiently and it allows the simulation of absorption and diffuse reflections. However, because

of the finite number of rays emitted from the source, the accuracy of the simulation result cannot be guaranteed. Ray tracing can be modified to trace beams instead of rays which enables the use of point receivers [25, 26]. This approach can be called cone tracing as opposed to the more precise beamtracing which is explained later [14].

Image Source Model

Image source model (ISM) is based on mirroring the sources with respect to the walls [27]. These mirrored *image sources* represent individual reflections and behave similarly to real sources (see Figure 4). The image sources themselves can further be mirrored with respect to the planes defined by other walls. Such second order sources represent reflections of reflections and the model can be generalized to any order. The generalization to the third dimension is straightforward. Although ISM is based on geometrical acoustics assumptions, Allen and Berkeley [27] proved that it offers the exact solution to the wave equation in rectangular rooms with perfectly rigid walls.

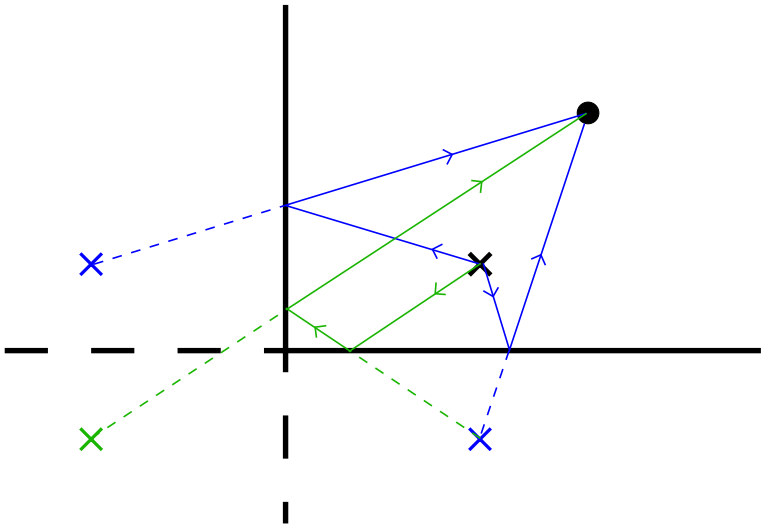

Figure 4: The image source principle. The solid colored lines represent the reflection paths from the source (black x) to the receiver (dot). The equivalent image sources (blue x) are formed by mirroring the source with respect to the walls. The second order image source (green x) is formed by mirroring an image source with respect to the plane defined by the wall.

Borish [28] expands the image source calculation to arbitrary polyhedric rooms. The calculation can be done by using vectors to the sources and normal vectors of the planes corresponding to the walls. The number of image sources increases quickly as each image source is mirrored with respect to each wall. Depending on the room geometry, the mirroring process may create invalid, invisible or, in the case of non-convex geometries, obstructed image sources [28]. Checking the criteria for all these false sources significantly adds to the computational complexity.

In its basic form, ISM models only specular reflections. Solutions have been proposed for modeling absorption [29], edge diffraction [22, 23] and diffusion [30]. The signal of each image source can be filtered with a set of filters modeling the absorption on the path from the source to the receiver [29]. Magnitude responses of different filters can be derived from the frequency dependent absorption coefficients $\alpha(f)$ with the following formula:

$$|H(f)| = \sqrt{1 - \alpha(f)}. \tag{12}$$

Various filters are then cascaded depending on the surfaces and materials on the path from the specific image source. The knowledge about the surfaces on the reflection path can be retrieved from the image source model. Path length -dependent air absorption can be modeled with an additional filter [29].

Edge diffraction is important for listener positions shadowed by obstructions but can be audible also in non-shadowed zones [31]. Svensson et al. [22] derive an analytical solution for impulse response calculation of finite-length edges. Using this model, edge diffraction can be modeled with diffraction sources. The diffraction from an edge is not point-like but diffraction sources offer efficient representations of the diffraction. Point source representations can also be justified by the fact that most energy is concentrated at the point of the edge that gives the shortest path between the source and receiver [31]. Pulkki et al. [23] present an implementation where the image source model is generalized to include edge diffraction using image sources for all the diffractive edges.

Simplified modeling of diffuse reflections can also be done with a digital filter structure as suggested in [30]. Such an approach fails to model the spatial image of the diffuse reflection and the diffuse energy spreading to other parts of the room eventually reaching the listener from other directions. Another approach for diffuse reflection modeling is a combination of several modeling schemes where specular reflections are modeled with an ISM and diffuse reflections with other approaches such as radiosity [32].

Beamtracing can be seen as a more efficient approach to the image source model. It treats the emitted sound as beams that lead from the point source to each polygonal surface in the room around it [33]. The beams are reflected using image sources as their apexes. The reflected beams are split into narrower beams, one for every surface they reach. This creates a so called beam tree. Beamtracing does not suffer from the problems caused by finite sampling of the space in ray tracing and is more efficient for complex room shapes than the ISM since the validity and visibility checks are not required. The geometrical computations are more complex than in the other methods but the system can be accelerated by preprocessing the room geometry and storing the beam tree from which the propagation paths can be found in run-time [33].

Element-Based Methods

Alternatives to the previously mentioned geometrical methods, which observe sound rays, are geometrical acoustic methods relying on computations of elements in the

room boundaries [14]. As many other geometrical acoustics modeling approaches, radiosity is a method inherited from computer graphics. The idea is to divide surfaces into patches and compute the energy transfer between different patches. The transfers between the patches are calculated step by step leading to time-variant energy responses at the patches. In the end, energy from the patches is gathered in the receiver position. Diffuse reflections are supported by definition since the energy transfers between the patches are not linked to specular reflections in any way.

Acoustic radiance transfer (ART) is another method adopted from computer graphics. ART is based on the room acoustic rendering equation, the acoustic counterpart of the rendering equation used in computer graphics [34]. Reflections are characterized by the bidirectional reflectance distribution functions (BRDF) at points on the surfaces. The rendering equation is built using the BRDFs and geometrical relations of boundary points. Dividing the surfaces into patches in a similar way as in the radiosity method, the energy propagation can be calculated using the acoustic rendering equation at the patches. ART can thus be seen as a generalization of the radiosity method to arbitrary reflection types. Siltanen [34] shows that other geometrical acoustic modeling techniques, such as ISM and ray-tracing can similarly be seen as special cases of ART.

3.2 Late Reverberation Modeling

Late reverberation of a room behaves approximately as exponentially decaying noise and the reverberant field in most rooms is nearly diffuse. These assumptions make it possible to take a perceptual approach instead of a physical modeling approach for the late field modeling. Digital reverberation effects have a long history [35, 36, 37, 38, 39]. A good overview on various methods can be found in [40].

Schroeder [35] started the use of allpass filters and comb filters as the basic building blocks of digital reverberator structures. His cascaded allpass filters offered an efficient way to create dense impulse responses with no or very little coloration. The parallel comb filters were used to model room modes and allpass filters created the necessary build-up of echo density. Schroder suggested several structures combining these efficient buildings blocks. Moorer [36] extended Schroeder's structures with lossy comb filters and connected the previously experimental results to more general room acoustic theory. Moorer's comb filters include a simple lowpass IIR in the feedback path damping the high frequencies of the reverberation similarly to typical absorption in real rooms.

D'attorro [38] suggested further design aspects for allpass filter networks and presented a structure that performs perceptually well but is based mainly on fine-tuning of specific parameters. Smith [37] introduced waveguides as a general audio signal processing tool and showed that networks of waveguides are powerful for reverberation modeling. They have not, however, gained popularity due simpler structures available. Smith and Rocchesso [41] have shown that waveguide meshes and feedback delay networks (described below) are isomorphic in certain cases.

Stautner and Puckett [42] introduced the idea of feedback delay networks (FDN) which Jot and Chaigne [39] generalized further. An FDN consists of parallel comb

filters like Schroeder's original designs but instead of keeping the comb filters as separate structures their feedback paths are combined through a matrix (see Figure 5). Depending on the matrix coefficients, the feedback paths are mixed together. This creates a significantly higher echo density since delay lines of different lengths are in a way connected in series. Following the idea of Moorer's loss filters inside the comb filter structures, filters can be added to the feedback paths to control the reverberation decay. The loss filters can be designed to create a desired frequency-dependent reverberation time by taking into account the length of the preceding delay line. Jot [39] suggests a tone corrector filter to be placed at the output of the FDN. Designing the tone corrector's frequency response inversely proportional to the frequency-dependent reverberation time equalizes the output and enables the frequency response to be adjusted with a separate filter.

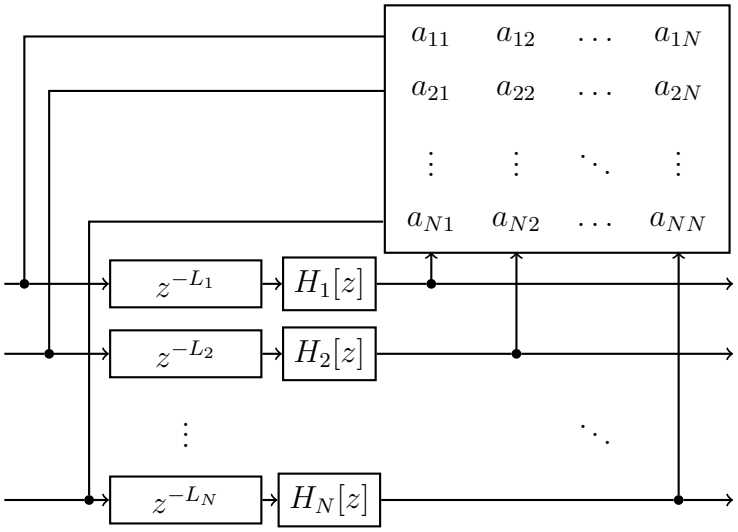

Figure 5: The structure of a basic feedback delay network consisting of delay lines, feedback loss filters and a feedback matrix.

In the FDN, as in other digital reverberator structures, the delay line lengths are usually chosen to be mutually prime. This avoids the echoes being piled up on the same samples which reduces patterns and hence frequency peaks and offers a smooth reverberation [36]. The feedback matrix should be unitary so that it does not have an effect on the frequency response of the reverberator [43]. In order to achieve as high echo density as possible, maximally diffusive matrices are typically used. These include Hadamard matrices and circulant matrices based on Galois sequences [44]. Sparse unitary matrices can be used to reduce computational complexity while keeping diffusiveness high [45]. In order to increase the echo density, Väänänen et al. [46] added allpass filters in the feedback paths of a simplified FDN where the feedback matrix calculations are avoided by simply summing the channels and feeding the sum back to the input.

3.3 Room Modeling System Description

The existing room modeling system for which automatic parameterization tools were developed in this work consists of an image source model for the early reflections and a feedback delay network for the late field. In addition to the basic ISM, the existing system models source directivity and material absorption with one biquad for each image source. The model is given the locations of the sources, the room geometry and material types of each wall. The parameterization system implemented in this work was developed for constant source and listener positions. In this case, the room geometry does not have to be estimated. The locations of the sources and their image sources is sufficient. In addition, the propagation path filters for each image source should be parameterized.

The FDN consists of 12 delay lines, a Hadamard matrix and feedback filters. There is also an equalization filter modifying the overall frequency response of the FDN. The parameters of the FDN thus include the frequency-dependent reverberation times, starting time of the FDN with respect to the direct sound and the overall frequency response of the late reverberation.

4 Room Acoustic Measurements

Room acoustic measurements are typically performed in order to calculate various room acoustic parameters, to examine specific acoustic phenomena occurring in a room or to reproduce the acoustics of the room. In this work, room acoustic measurements are used to obtain the information needed for the parameterization of the room model and thus for reproduction of the room acoustics.

4.1 Room Impulse Response Measurements

As explained in Chapter 2, the room impulse response describes the propagation path between a source and a receiver. In general, an impulse response defines a linear time-invariant system and the Fourier transform of the impulse response is the transfer function of the system. Impulse response measurements are the most common acoustic measurements because impulse responses are useful in assessing the properties of rooms, loudspeakers and other acoustic systems.

Müller and Massarani [47] give a good overview and comparison of various impulse response measurement methods. The simplest method to get an approximation of an RIR is to excite the room with an impulse-like sound and record it. Balloon-pops and start pistols are easy ways to generate approximate impulses. Using a loudspeaker connected to the measurement system to reproduce the impulse allows better control of the excitation signal and allows the propagation time of the sound to be measured.

Modern impulse response measurement systems typically use either pseudo-random signals, such as maximum length sequences (MLS), or swept sines [48]. In these methods, the excitation signal is transmitted, recorded, and the recorded signal is deconvolved with the original excitation signal in order to retrieve the impulse response. The main benefit of using logarithmic sweeps instead of MLS is strong rejection of harmonic distortion and, consequently, a high signal-to-noise ratio (SNR) [47].

4.2 Reverberation Time

Architectural acoustic research has spawned a number of statistical measures of room acoustics that can be calculated from an RIR or a number of RIRs measured with a specific microphone combination. Most of the measures are mainly used for describing concert halls and auditoriums and are rough statistical measures related to perceptually important aspects of concert hall acoustics, such as the reverberation decay time, strength of the reverberation, relative frequency content of the reverberation and the correlation of the sound arriving at the left and right ears. Reverberation time is the single most important measure and is widely used because it is perceptually important and has been extensively studied theoretically (see Chapter 2).

Standardized procedures for reverberation time measurements in performance spaces and in normal rooms are described in ISO standards 3382-1 [49] and 3382-2

[50], respectively. The standards suggest using omnidirectional sources and receivers in six source-microphone position combinations. Frequency-dependent reverberation times are analyzed from impulse responses measured at the position combinations and averaged over the positions.

As was mentioned in Chapter 2, the reverberation time RT_{60} of the room is defined as the time it takes for interrupted constant signal to decay 60 dB [50]. This decay is commonly calculated from an RIR measured with one of the methods described above. Schroeder [51] showed that taking an ensemble average over infinitely many squared decay curves of interrupted noise is equal to integrating backwards a single impulse response measured at the same transmitter and receiver positions. The backwards integration leads to the energy decay curve (EDC):

$$s[n] = \sum_{i=n}^{\infty} r^2[i], \tag{13}$$

where $r[n]$ is the measured RIR.

A typical decay curve of a room is approximately exponential and thus linear on a logarithmic scale. Reverberation time is theoretically constant throughout the decay since it is the time constant associated with a decay of 60 dB. When a smaller dynamic range than 60 dB is available, reverberation time is typically measured from a 20 or 30 dB interval and multiplied by the corresponding factor (3 or 2) to obtain the 60 dB time interval. In standard reverberation time calculations, the estimation is done starting from 5 dB below the peak of the decay curve. The beginning of the decay is generally not used since it might not behave according to the exponential decay depending on the source distance and room volume. The beginning of the response is used when calculating another measure called the early decay time (EDT). It is the time constant of 60 dB decay as well but calculated using the early decay from 0 dB to -10 dB.

In all real rooms and microphones, there is also some background noise constantly present. The noise floor causes offset from the exponential decay and causes errors to reverberation time estimation. The solution suggested in the standard [49] is to estimate the level of noise from the end of the squared impulse response and the line representing the exponential decay on a logarithmic scale and to find where these two lines cross. The found time limit T_{lim} is then used instead of infinity in the EDC calculation. In both RT_{60} and EDT calculation, a line is typically fitted to the decay in order to get a more stable estimate [49]. Reverberation times are presented in octave or third-octave bands. The RIRs are filtered to these bands and the EDC calculation and curve fitting is done separately in each band.

Reverberation time is used to describe the acoustics of an entire room and thus spatial averaging is usually required. For measurements in performance spaces, the recommended minimum number of source and receiver locations are 2 and 3, respectively [49]. For precision measurements in normal rooms, the minimum numbers are the same but the recommended number of source-receiver combinations is 12 [50]. If the reverberation needs to be described only for a single source-receiver position combination, such as in the case of auralization with static source and listener position, no spatial averaging is required.

The traditional combination of bandpass filtering and EDC calculation has a limited frequency resolution and precision depending on the properties of the filterbank [7]. Jot et al. [7] expanded Schroeder's EDC to the frequency domain introducing the energy decay relief (EDR). The idea is to use short-time Fourier transform (STFT) or another time-frequency representation of the RIR and calculate the EDC in each frequency band.

Jot et. al [7] propose an iterative approach for finding the frequency dependent reverberation time $RT_{60}(f)$, the noise floor power spectrum $P_{\mathrm{n}}(f)$ and initial power spectrum of the decay $P(f)$. The analysis begins with the computation of the spectrogram using STFT and the computation of the EDR by backwards integrating the spectrogram. Initial estimates of the parameters can then be calculated from the EDR. The iterative process goes as follows [7]:

I Estimate $T_{\mathrm{lim}}(f)$ as the time when the ideal exponential decay attenuates below $P_{\mathrm{n}}(f)$.

II Compute $P_{\mathrm{n}}(f)$ by averaging over the spectrogram from $T_{lim}(f)$ to the end of the response from which the exponential decay has been removed.

III Subtract the $P_{\mathrm{n}}(f)$ from the spectrogram and calculate the EDC.

IV Estimate $T_{\mathrm{lim}}(f)$ using a linear fit to the EDC in the logarithmic domain.

These steps are repeated until the errors in the linear fit are small enough.

4.3 Normalized Echo Density

Echo density has been long considered as an important factor in the quality of reverberation effects [35]. Abel and Huang [52] developed a measure of echo density which can be calculated from an RIR. For a zero-mean RIR, the normalized echo density (NED), also called the echo density profile, is defined as the fraction of impulse response taps in a sliding RIR window lying outside the window standard deviation:

$$\eta[n] = \frac{1}{\mathrm{ercf}(1/\sqrt{2})} \sum_{l=n-\delta}^{n+\delta} w[l]\mathbf{1}\{|r[l] > \sigma|\}, \tag{14}$$

where $\mathrm{ercf}(1/\sqrt{2}) = 0.3173$ is the expected fraction of samples lying outside the standard deviation from the mean for a Gaussian distribution, $\mathbf{1}\{.\}$ is the indicator function returning one for true arguments and zero for false, $r[l]$ is the room impulse response, $w[l]$ is a weighting function attenuating the impulse response on the sides of the sliding window and normalized to unit sum $\sum_l w[l] = 1$ and

$$\sigma = \sqrt{\sum_{l=k-\delta}^{k+\delta} w[l]r^2[l]} \tag{15}$$

is the standard deviation of the impulse response within the window.

Typically NED of an RIR starts from zero and approaches one as individual reflections start to arrive more densely and the RIR approaches Gaussian noise. Since NED is described as a fraction, the value remains around one for the rest of the impulse response. The slope of the rising NED curve depends on the room properties. The point where a value close to one is reached can be assumed to be the beginning of the statistical late field. In other words, NED can be used as an empirical indicator for the mixing time. Due to natural fluctuations in the NEDs of real rooms, Abel and Huang [52] suggested the mixing time NED threshold to be $1 - \sigma_{\text{late}}$, where σ_{late} is the standard deviation of the late field NED.

The length and type of the analysis window have a strong effect on the form of the echo density profile. A short window leads to a quickly varying NED since it is likely that on two consecutive time instances a short window includes a reflection in the first time instant and includes no reflections in the second one. Longer windows smooth the response. Abel and Huang [52] recommend a window length of 20-30 ms as a compromise between smoothness and fast enough response time to changes. More smoothness to the echo density profile can be achieved by using a window function that attenuates the taps at the edges of the time window thus introducing strong reflections gradually to the NED. Abel and Huang used a Hanning window [52].

NED has proven to be insensitive to decay rate, equalization, level and sampling rate [52]. It can also be connected to the absolute echo density, i.e. the number of echoes per second, which is the typical way to represent echo density [53]. Huang et al. [54] also proved that the NED is a perceptually important measure connecting the diffuseness, or density of echoes, to perceptual textures of the different phases of the reverberation.

4.4 Mixing Time Estimation

In addition to NED, several other measures have been used to define empirically the mixing time, the starting point of the diffuse late field. Lindau et al. [10] compared different empirical measures useful for finding the mixing time of a room. NED was compared to three other methods that can be calculated from impulse responses: Steward and Sandler's [55] kurtosis, Hidaka's [11] method and the matching pursuit -based method created by Defrance et al. [56].

Kurtosis is the fourth order zero-lag cumulant of a zero-mean process defined as [55]

$$\gamma_4 = \frac{\mathrm{E}(x-\gamma)^4}{\sigma^4} - 3, \tag{16}$$

where E() is the expectation operator, γ is the mean and σ is the standard deviation of the process. Typically this measure approaches zero towards the end of an RIR because of its increasingly Gaussian nature. Lindau et al. [10] estimate the mixing time as the time when the kurtosis calculated in windows similarly to NED reaches zero. Hidaka's [11] method is based on integration of the time-frequency energy distribution in a similar sense as in the calculation of EDR. The energy integral is used in the computation of Pearson's product-moment correlation which is assumed

to define the mixing time as it gets small enough. The method of Defrance et al. [56] uses a decomposition computation called matching pursuit, which is described in Chapter 6. In their method, the occurrence times of reflections are searched using the similarity of the direct sound to the rest of the impulse response. The found arrival times of the reflections are used to find a point in time where the distance between two reflections is shorter than the so called equivalent duration of the impulse calculated from the direct sound [56].

Perceptual tests show NED to be the most reliable measure for estimating the mixing time [10]. Using the threshold value of $1 - \sigma_{\text{late}}$ as suggested by Abel and Huang [52] does not show improved prediction power. Lindau et al. [10] derive a regression formulas for the NED-based mixing time $t_{\text{mixingNED}}$ based on the perceptual tests:

$$t_{\text{mixing50\%}} = 0.8 t_{\text{mixingNED}} - 8 \tag{17}$$

and

$$t_{\text{mixing95\%}} = 1.8 t_{\text{mixingNED}} - 38, \tag{18}$$

where $t_{\text{mixing50\%}}$ is based on the mean perceptual mixing times and $t_{\text{mixing95\%}}$ on the 95% percentile of the perceptual mixing time values.

5 Acoustic Source Localization

Source localization is used in a large variety of fields ranging from navigation and telecommunications to seismology and military surveillance. Localization based on acoustic signals can be used for instance in underwater acoustics, teleconferencing systems and hands-free speech acquisition. Due to the large variety of applications, plenty of research has been done in the area. Source localization is based on using an array of sensors, microphones in acoustic source localization, and combining or comparing their input signals to retrieve the source locations or directions. Source localization can be active or passive, active methods sending and receiving energy and passive methods only receiving energy [57].

Usually the microphone arrays used for localization are composed of omnidirectional microphones placed in a known formation at predetermined distances from each other. For a single point source, each microphone receives the same signal but at different times and amplitudes depending on the source location and the array geometry. The source localization algorithms differ in the way they process the signals to retrieve a location estimate. Many algorithms can be generalized to various, if not arbitrary, array geometries. Choosing the array geometry has, however, a strong effect on the performance and limitations of the localization.

In this chapter, a few important categories of acoustic source localization methods are presented. Most emphasis will be placed on the methods used in this work and alternative methods suitable for localizing reflections.

5.1 Beamforming

Using microphone arrays and simple processing of the signals recorded at the different microphones, it is possible to direct the pick-up pattern of the microphone array to desired directions. This process is called beamforming and is the basis for several source localization techniques. In the simplest case of beamforming the signals of the microphones are simply delayed depending on the array geometry and the desired beam direction and summed together. This is called the delay-and-sum beamformer. Different combinations of delays represent the directional responses which is due to the fact that with properly chosen delays signals from certain directions get summed in phase and others out of phase fortifying and attenuating the signals, respectively. Replacing the delays with FIR filters in the so called filter-and-sum beamformer enables the frequency-dependency of the beam patterns to be controlled. It is common to use adaptive beamforming techniques, such as the Capon's beamformer (also known as minimum variance distortionless response beamformer, MVDR) to minimize the effects of noise coming from directions other than the look direction. [58]

The direction of arrival (DOA) can be analyzed as the direction which maximizes the power of the beamformer [57]. This is called the steered response power (SRP) approach. An iterative approach, such as the incremental multi-parameter (IMP) algorithm [59, 60, 61], can be used to localize several sources. In the basic form, IMP is based on finding the direction of the strongest signal and nulling out that

direction in the response. This procedure is repeated until the response has no clearly detectable arrival directions left.

Although some methods analyze the curvature of the wavefront, most beamforming techniques assume far field sources. The arriving sound waves are thus assumed to be plane waves. No distance data is therefore available from the array processing directly. If the transmission time of sound signal is known, as in impulse response measurements described in the previous chapter, the distance can be estimated and coordinates with respect to the array calculated based on the DOA and the distance.

Beamforming is band-limited because it is based on the detection of phase differences of the signals at a finite number of microphones. Spatial aliasing will occur for frequencies higher than $f_{\max} = c/2d$ where c is the speed of sound and d is the microphone spacing [58]. The spacing of the sensors must hence be small enough in order to avoid localization ambiguity due to spatial aliasing in the observed frequency band. Very small microphone spacing relative to the wavelength of sound, on the other hand, can reduce the resolution of the localization [62].

More developed beamforming-based source localization methods commonly use the spatial covariance matrix [63] which for zero-mean signals is defined as

$$\mathbf{R} = E[\mathbf{x}(t)\mathbf{x}^{\mathrm{H}}(t)] \tag{19}$$

where $E[]$ is the expected value operator, H denotes the Hermitian transpose and $\mathbf{x}(t)$ is the vector of signal values at the sensors of the array at time t. In practice, the spatial covariance matrix is estimated with the sample covariance matrix

$$\hat{\mathbf{R}} = \frac{1}{N}\sum_{n=0}^{L-1}\mathbf{x}[n]\mathbf{x}^{\mathrm{H}}[n]. \tag{20}$$

In subspace-based methods, the spectral decomposition of the spatial covariance matrix is calculated and the received signal is divided into noise and signal subspaces. MUSIC (multiple signal classification) [64] is a classic subspace-based localization method. Its main limitation is that it allows only incoherent signals to be separated and localized. The reflections in a room acoustic environment, however, are typically coherent. Extensions to MUSIC include the MIN-NORM algorithm [65] which can be seen as a weighted version of MUSIC [63]. ASPECT (adaptive signal parameter estimation and classification technique) is an alternative to MUSIC capable of separating also coherent signals. Other advanced spectral estimation methods include Root-MUSIC and ESPRIT both of which can only be used with uniform linear arrays [63].

5.2 Time Delay Estimation -Based Methods

5.2.1 Time Difference of Arrival Estimation

Reliable time difference of arrival (TDOA) estimation is at the core of several source localization algorithms. TDOA estimation is a preliminary phase of source localization where the differences between the times of arrival at the sensors in an array

are estimated. Typically in TDOA estimation, a measure of similarity between the signals of various sensors is used. Several options exist for the similarity function, two popular ones being the generalized cross-correlation (GCC) [66] and average magnitude difference function (AMDF) or MAMDF [67].

For two discrete-time signals $x[n]$ and $y[n]$ cross-correlation is defined as

$$R_{xy}[\tau] = \sum_{n=0}^{L-\tau-1} x[n]y[n+\tau], \tag{21}$$

where τ is the delay between the signals and L is the length of the signals. The Fourier transform of cross-correlation is called the cross power spectral density $G_{xy}[k]$. It can also be calculated using the Fourier transformed versions $X[k]$ and $Y[k]$ of the signals. The Generalized Cross-Correlation is then [66]:

$$G_{xy}[k] = W[k]X^*[k]Y[k] \tag{22}$$

where $W[k]$ is the weighting function. Popular weighting functions include cross-correlation (CC), phase transform (PHAT), smoothed coherence transform (SCOT) and maximum likelihood (ML) which are listed in the Table 1 [66].

Table 1: Different weighting functions for GCC [66]. Maximum likelihood (ML) weighting uses the coherence function $\gamma_{xy}[k] = \frac{G_{xy}[k]}{\sqrt{G_{xx}[k]G_{yy}[k]}}$

Name	$W[k]$
CC	1
PHAT	$\frac{1}{\lvert G_{xy}[k]\rvert}$
SCOT	$\frac{1}{\sqrt{G_{xx}[k]G_{yy}[k]}}$
ML	$\frac{1}{\lvert G_{xy}[k]\rvert}\frac{\gamma_{xy}^2[k]}{1-\gamma_{xy}^2[k]}$

The average magnitude difference function is [67]

$$R_{xy}^{\mathrm{AMDF}}[\tau] = \frac{1}{L}\sum_{k=0}^{L-\tau-1} \left|x[k] - y[k+\tau]\right|, \tag{23}$$

and similarly average magnitude sum function is

$$R_{xy}^{\mathrm{AMSF}}[\tau] = \frac{1}{L}\sum_{k=0}^{L-\tau-1} \left|x[k] + y[k+\tau]\right|. \tag{24}$$

The basic difference of these to GCC is the use of difference or sum instead of multiplication. AMDF and AMSF can be combined into MAMDF since they do not correlate [57]:

$$R_{xy}^{\mathrm{MAMDF}}[\tau] = \frac{R_{xy}^{\mathrm{AMDF}}[\tau]}{R_{xy}^{\mathrm{AMSF}}[\tau] + \epsilon}, \tag{25}$$

where $\epsilon > 0$ is a small number.

The time delay estimate is found by maximizing GCC or minimizing MAMDF. The accuracy of the time delays is limited by the sampling frequency and therefore interpolation around the peak of the similarity function is usually applied. Typically, a parabola is fitted to the three closest points giving the time delay estimate as the location of the parabola's maximum [68]. Exponential interpolation can lead to more accurate and robust estimation [69]. In the multiple source scenario there are several maxima or several minima and their association with specific sources needs to be determined.

5.2.2 Source Localization Based on Time Delays

Assuming that the speed of sound is known, the TDOA estimates are proportional to the distance differences from the sensors to the source, also known as range differences (see Figure 6). Methods based on TDOAs often assume that the source is in the near-field and thus the observed wavefront is spherical. Hence, range difference estimates define hyperboloids in the 3D space for each range difference estimate. In theory, several hyperboloids can be calculated from a number of range difference measurements and the source could be localized in the intersection of these hyperboloids. In practice, there typically is no common intersection due to measurement noise and thus the location must be estimated. The source localization problem based on TDOAs can be solved in closed-form, iteratively or in a sequential Bayesian framework [57].

The various closed-form solutions are explained e.g. in [57, 70, 71]. Typically these solutions rely on minimizing a localization error in the least squares sense. Different approaches have been taken to overcome the lack of knowledge about the distance to the source. Only the simplified case using a known distance to the source is presented here because in the context of this thesis the distance can be estimated from the impulse response. Most estimators assume the speed of sound to be known. In some cases, it is possible to do joint estimation of the speed of sound [72] but in this thesis relatively constant room temperatures are assumed and thus a generic value for the speed of sound is used.

The formulation of the least squares problem [73] starts by setting one of the microphones as the origin of the coordinate system. Let it be called the reference microphone. TDOA estimation gives the time differences $\Delta\tau_i$ $(i = 2...N)$ between the reference microphone $i = 1$ and the other $N-1$ microphones. The corresponding range difference estimates are

$$d_i = \Delta\tau_i c, \tag{26}$$

where c is the speed of sound. Since the reference microphone is at the origin, the range differences are by definition

$$d_i = ||\mathbf{x}_i - \mathbf{x}_\mathrm{s}|| - ||\mathbf{x}_\mathrm{s}||\,, \tag{27}$$

where $\mathbf{x}_i$ is a vector specifying the coordinates of the ith microphone and $\mathbf{x}_\mathrm{s}$ is the unknown location of the source. Denoting the distances from the origin by $R_i = ||\mathbf{x}_i||$

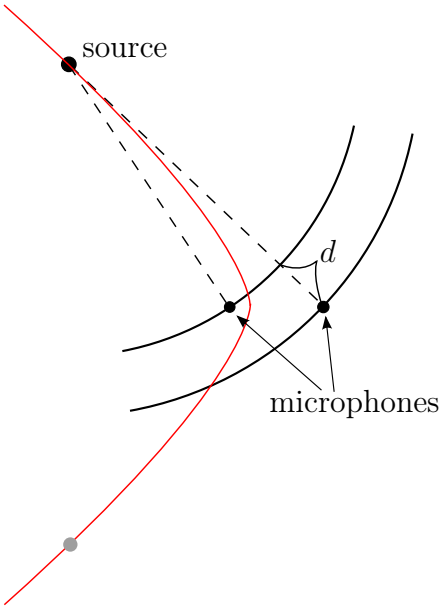

Figure 6: TDOA-based localization. The range difference d can be calculated from the TDOA between the two microphones. Based on the TDOA, the source is localized on the hyperbola (red). If knowledge of propagation path length is available (based on time of arrival estimation), the source location is known to be either at the actual source position or at the grey dot. Range difference measurements with respect to the other coordinate axis is thus required. In the 3D case, the hyperbola is replaced by a hyperboloid and measurements with respect to all three coordinate axes are required.

and reformulating (27) we get

$$(d_i + R_\mathrm{s})^2 = ||x_i - x_\mathrm{s}||^2 \tag{28}$$

and then

$$d_i^2 + 2d_i R_\mathrm{s} + R_\mathrm{s}^2 = R_i^2 - 2x_i^\mathrm{T} x_\mathrm{s} + R_\mathrm{s}^2 \tag{29}$$

Because of the measurement and estimation errors, Equation (28) does not hold exactly and we introduce an error measure

$$\epsilon_i = R_i^2 - d_i^2 - 2d_i R_\mathrm{s} - 2x_i^\mathrm{T} x_\mathrm{s}. \tag{30}$$

Now the error can be minimized in the least squares (LS) sense using $N-1$ equations in the matrix form:

$$\epsilon = \delta - 2R_\mathrm{s}\mathbf{d} - 2\mathbf{S}\mathbf{x}_\mathrm{s} \tag{31}$$

where δ is a column vector with elements $\delta_j = R_{j+1}^2 - d_{j+1}^2, j = 1...N-1$, d is a column vector of range differences and $\mathbf{S}$ is a matrix with the microphone coordinate vectors as its rows. The LS solution for the source location is

$$\mathbf{x_s} = \frac{1}{2}(\mathbf{S}^\mathrm{T}\mathbf{S})^{-1}\mathbf{S}^\mathrm{T}\left(\delta - 2R_\mathrm{s}\mathbf{d}\right) \tag{32}$$

For the solution above, $N-1$ time differences were used. The data available offers $N(N-1)/2$ time delay and range difference estimates. The other estimates are redundant data but they can be used to get better performance in the presence of noise especially using methods fusing the similarity functions of the microphone channels [74]. One such fusion method is the steered response power using phase transform (SRP-PHAT) [75] in which the similarity functions of several microphone pairs are summed and localization is done based on the combined data. The closed-form localization formulation can also be extended to use redundant estimates [57].

5.3 Intensity Vector -Based Methods

Intensity vectors offer another approach to the DOA estimation. Sound intensity is a vector quantity with a strength and a direction. Estimating intensity vectors can thus give the DOA at a given time in a given frequency band [76]. Intensity vector -based direction estimation methods are compared in [77].

Sound intensity can be calculated when the pressure and particle velocity at a point in space is known. Most microphones measure sound pressure and the simple averaging can be used to estimate the pressure between microphones in an array. Particle velocity in one coordinate axis can be estimated using the difference between the measurements at these points and three-dimensional intensity vectors are estimated using three microphone pairs, one for each coordinate axis. Intensity vectors can also be estimated from B-format signals and there is also specific equipment for direct particle velocity measurement [76]. After the estimation, some bias compensation is typically required [77].

There are two groups of methods for direction estimation using a group of intensity vectors on different frequency bands: direct and mixture models [77]. Direct methods typically use some kind of averaging over the intensity vectors, such as circular mean or circular median, to find the direction whereas mixture models fit a mixture of distributions to the angle distribution of the intensity vectors. Mixture models are generally more reliable and provide good rejection for additive noise.

5.4 Previous Studies on the Localization of Reflections

Due to the dominating applications, many algorithms are developed to suppress the effects of the room (see e.g.[71]). Korhonen [72], on the other hand, introduces methods using the reflections to enhance the localization of the real sources. In this work, the aim is to localize the individual reflections.

The localization of reflective surfaces has been studied earlier from several perspectives. Günel [78] and Aprea et al. [79] use a method which requires the measurement microphone to be moved around the loudspeaker exciting the room. Kuster [80] uses inverse extrapolation of the Kirchoff-Helmholtz and Rayleigh integrals to create acoustic images of reflective surfaces. Tervo and Korhonen [81] suggest a method which uses source localization and inverse mapping of the multipath propagation.

In a technique called spatial impulse response rendering (SIRR) [82, 76], impulse responses measured with a microphone array are analyzed for reproduction of the room acoustics. DOAs are calculated in frequency bins of short time windows using intensity vectors. However, the actual reflection source locations are not analyzed further.

Large spherical microphone arrays have been used to visually inspect reflections [83, 84]. Beamforming with a large number of microphones along a sphere makes it possible to inspect the directions of arrival of sound waves in real time. Integrated video cameras have been used to show the directions of sources and reflections overlaid to image of the microphone array's surroundings.

The problem of localizing individual reflections has previously been tackled in [85, 86, 62]. Van Lancker [85] uses time delay estimation -based localization with an eight-capsule cubical microphone array. Tervo et al. [86] compare three different source localization methods for the localization of reflections using impulse responses measured with a spherical microphone array. In their experiments, cross correlation -based localization works better than intensity vector -based localization. Roper [62] uses a large 2-D microphone array which is a combination of a line array and circular array. Localization is done in two dimensions using IMP.

In a recent study, published after the work presented in this thesis had been finished, Tervo et al. [87] make a comprehensive investigation of reflections localization methods. Time of arrival and time difference of arrival approaches are presented and combined. The approach for reflection localization taken in this work and presented in the following chapter uses time of arrival information and time difference of arrival information in a sequential matter where reflections are first detected and extracted and the actual localization is then done using time difference of arrival -based localization.

6 Analysis of Direct Sound and Early Reflections in the Implemented System

As mentioned in Section 3.3, the starting point for the work in this thesis was an existing auralization system. The parameterization process of the direct sound processing and image source model in the existing system requires specifying the source and wall positions with respect to the listener as well as wall materials in terms of their frequency dependent absorption. The image source locations are calculated from the source and wall positions offline and used as sources in run-time processing. In the case of static source and receiver locations, an automatic parameterization system can rely on directly localizing the image sources instead of the walls. In addition to image source localization, the parameterization of the ISM requires approximating the frequency-dependent absorption effects on the path from each image source.

The starting point of the analysis described below is an impulse response measurement with a tetrahedral microphone array consisting of four omnidirectional microphone capsules (see Figure 7). This array configuration was settled on because of the small size and the minimum number of microphones able to span three dimensions. The tetrahedral form places the microphones evenly on surface of a sphere leaving open the possibility for using spherical beamforming techniques. The mounting was kept as small as possible to alter the omnidirectional behavior of the microphone capsules as little as possible.

Figure 7: The tetrahedral array used for room impulse response measurements.

The impulse response measurements are done from each source (typically loudspeakers of a surround system) to each of the four microphones. For this thesis, all measurements were made using logarithmic sweeps as excitation signals. The analysis system calculates measures of the noise floor from the end of the impulse

response and thus assumes that the measurements are sufficiently long. A sufficient length here is at least two times the maximum reverberation time. A robust and generalizable system cannot assume the loudspeaker impulse response to be known. This creates challenges in the analysis since the measured impulse responses are, in theory, convolutions of the room impulse responses and the speaker impulse responses.

The structure of the analysis process described in this chapter is depicted in Figure 8. First, the direct sound and the individual reflections in the impulse responses are separated using different filtering techniques and the matching pursuit algorithm. Sources, both real and image sources, are then localized using cross-correlation based TDOA estimation between the four microphones and closed-form least squares estimation to find the most likely location based on the TDOA estimates. For each image source, estimation of the spectral characteristics of the reflection are also estimated.

6.1 Direct Sound Extraction

The locations of the real sound sources are the most crucial parameters for the auralization and therefore direct sounds must be carefully extracted to achieve reliable location estimates. The extracted direct sounds are also used in the separation of the reflections. In addition, the extracted direct sounds and their frequency responses are necessary for the analysis of the spectral characteristics of the reflections as well as in the equalization of the late reverberation model. The extraction of direct sounds is thus the key for the success of the entire parameterization process.

In an ideal room impulse response, the direct sound is the first and strongest impulse and the reflections are delayed, attenuated and filtered versions of the direct sound. Due to inaccuracies in the measurement and deconvolution processes and especially because the source's impulse response is not known, the RIRs available in the analysis do not follow the ideal case. The direct sound in a measured impulse response is the convolution of an impulse and the impulse response of the loudspeaker with additional background noise. Moreover, the mounting of the capsules in the microphone array affects directivity at high frequencies and diffraction from the loudspeaker stand or other mounting and other objects around the loudspeaker spread the impulse response of the speaker. Consequently, the direct sounds and individual reflections may merge into each other making the separation of them difficult.

In a typical multiway loudspeaker, low frequencies get delayed with respect to the high frequencies. Heyser [88] explains this as a consequence of the inverse relation between the high-frequency cutoff and the additional time delay of the loudspeaker element which causes more delay to the response of the woofer than that of the tweeter. Since low frequencies arrive slightly later to the microphone than the high frequencies, most of the overlap of the direct sound and the reflections is at low frequencies. In addition, the low frequency fluctuations of the impulse response are harmful for the matching pursuit algorithm. Therefore, the impulse responses are highpass-filtered to remove the slow variations. In Figure 9, it can be seen how the

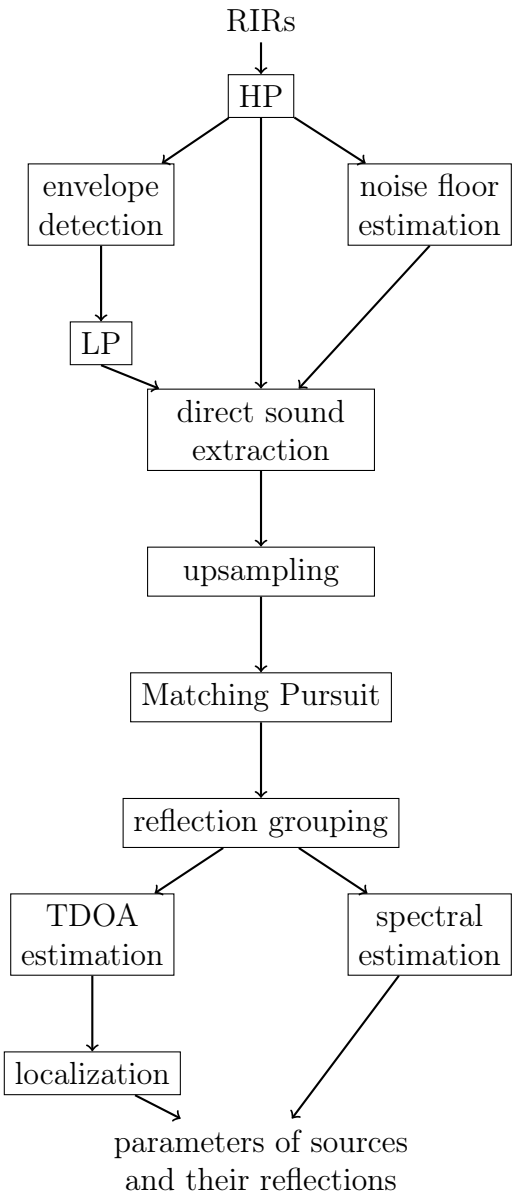

Figure 8: Block diagram of the analysis of direct sounds and early reflections. HP and LP refer to highpass and lowpass filtering, respectively.

highpass filtering removes the low frequency fluctuations in the impulse response during the following reflection. At a sampling rate of 44.1 kHz, an FIR highpass filter of the order 30 with 3 kHz cutoff frequency was used and the groupdelay of the linear phase filter was compensated in order to keep the timings in the impulse responses correct.

The starting point of the direct sound is estimated as the time instant when the absolute value of the impulse response first rises above the noise floor. The noise floor can be estimated using the root-mean-square (RMS) value of the ends of the impulse responses where the response is assumed to have decayed under the

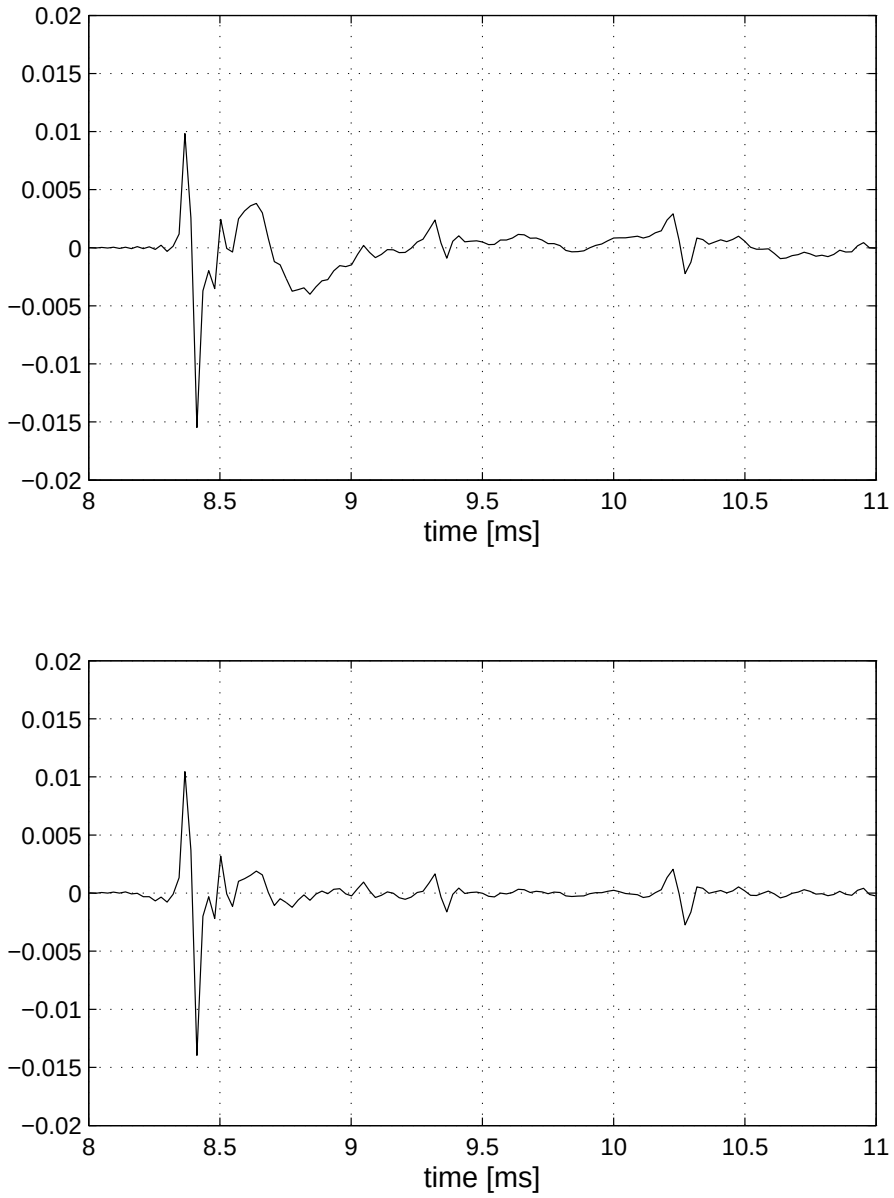

Figure 9: The effect of highpass filtering of the RIR. The delayed low frequencies of the direct sound are still visible during the first reflections. The highpass filtering makes the reflections around 9.3 and 10.25 ms stand out better and removes the amplitude offsets around them, making them easier to detect later.

noise floor. As mentioned in the beginning of the chapter, RIR measurements are assumed to be long enough to enable reliable estimations of the noise floor from the latter halves of the RIRs. In the parts of the impulse response consisting of only the background noise, the signal varies constantly around the RMS value. Thus, the direct sound is approximated to begin when the absolute amplitude of the signal exceeds a level high enough above the RMS noise. An experimentally derived 25 dB threshold was used in the implementation.

In order to find the ending point of the direct signal, the envelope of the impulse response is calculated. For an offline Matlab-based analysis system, Hilbert transform offers reliable envelope estimation. The envelope of the signal is given by the absolute value of the analytic signal [89]

$$x_{\text{env}}[n] = |x[n] + \mathrm{i}\mathcal{H}[x[n]]|\,, \tag{33}$$

where $\mathcal{H}[x[n]]$ is the Hilbert transform of the signal $x[n]$. An alternative way to calculate the envelope without the Hilbert transform is by taking the Fourier transform, zeroing the amplitudes of the negative frequencies, taking the inverse Fourier transform, calculating its absolute value and finally multiplying by two [89].

The peak of the envelope is used as an estimate of the arrival time of the direct sound. As seen with an example impulse response in Figure 10, the Hilbert envelope follows the fast fluctuations of the waveform. In order to make the direct sound show as a single "hill" in the envelope for end point estimation, the envelope is smoothed by lowpass-filtering. From this smoothed envelope shown in Figure 10 the end of the direct sound can be estimated as the first minimum after the first peak. This way the direct sound extraction end point is typically as late as possible but before the disturbing first reflection. An FIR filter with order 30 and cutoff frequency 1000 Hz (sampling rate 44.1 kHz) was used for the lowpass filtering of the envelope as it performed well for the impulse responses available. As before in the highpass filtering of the RIR, the group delay of the filter was compensated in order to keep the position of the envelope in the time domain stable. With the starting and ending points of the direct sound analyzed, it can be simply cut out of the impulse response, leaving the RIR with the direct sound part zeroed for further analysis of the reflections.

6.2 Reflection Extraction

Individual reflections must be extracted from the impulse response for separate TDOA estimation. Since the reflections are spread impulses, merge together and have smaller amplitudes in some microphones than others due to the shadowing of the mounting, the task is not simple. An iterative algorithm for finding best matches of the direct sound was thus chosen and is explained below.

Matching pursuit (MP) is an iterative algorithm used to find a representation of a signal using a dictionary of waveform atoms [90]. MP has been used previously as a non-linear deconvolution method for RIR measurements [91]. As the dictionary atoms, Gabor functions [90], damped sinusoids [92] and lowpassed impulses [91] have been used. Roper [62] uses a separately measured on-axis measurement of

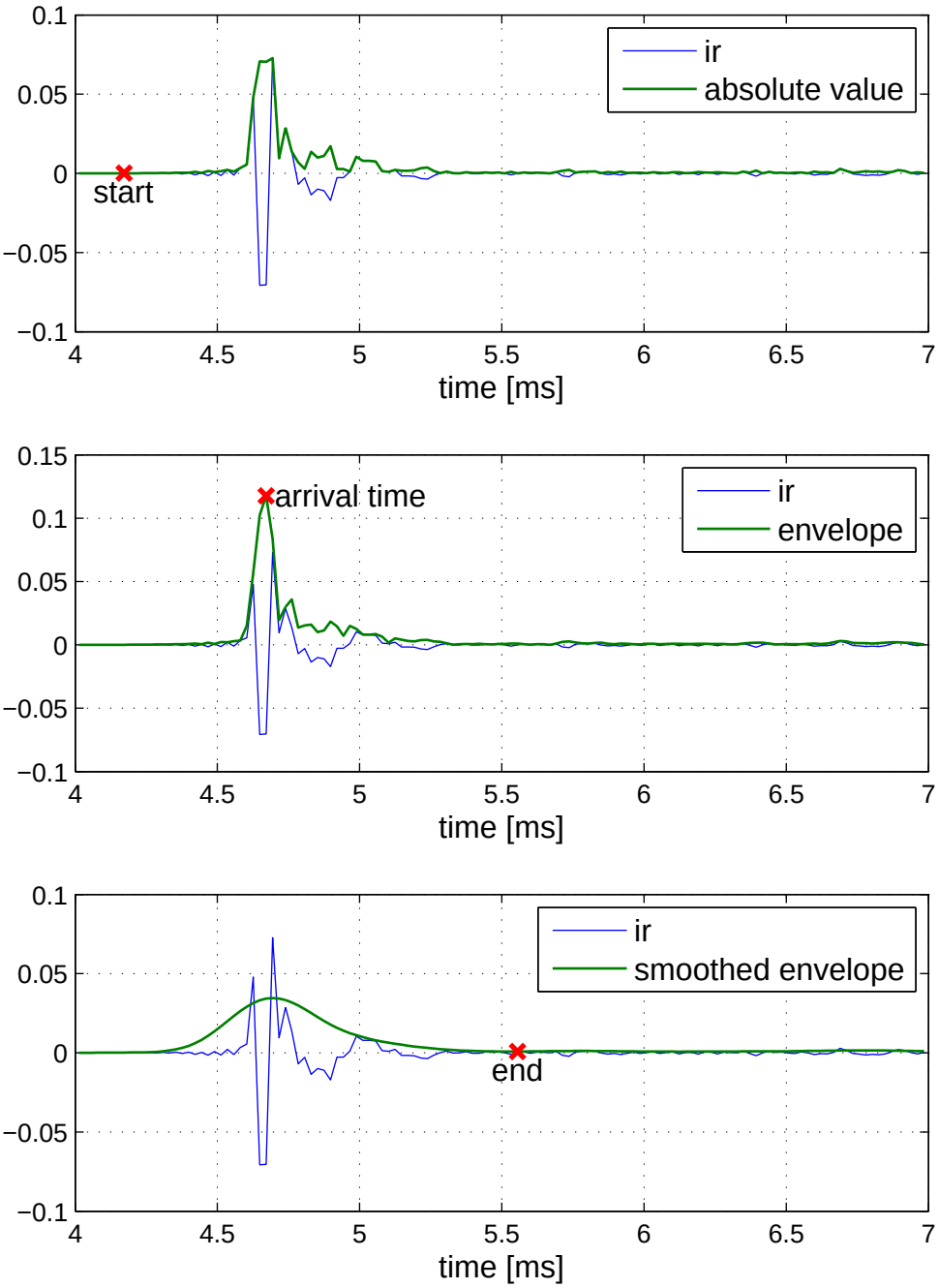

Figure 10: Direct sound extraction. Arrival time is estimated as the largest peak of the envelope, start time of the direct sound is the point when the absolute value of the RIR exceeds estimated noise floor and ending point where the smoothed envelope has its first minimum after the peak.

the loudspeaker impulse response as the only atom. Matching pursuit functions in his system as a combined non-linear deconvolution and reflected sound arrival time estimation method. Defrance [56] suggests running MP for several direct sound estimates with different lengths. The correct direct sound limits are assumed to be the ones that lead to the fewest iterations of the algorithm. The results of the MP run with these direct sound borders are used specifically for tracking how the echo density evolves with time.

On every round of the MP, the atoms are cross-correlated with the signal. The maximum value of all cross-correlations is found. The corresponding atom at a delay of the peak with the amplitude of the peak is added to the new representation of the signal and subtracted from the original signal. This is repeated until the signal residual is low enough or a predetermined number of iterations is reached.

It is assumed here that the general waveform of the loudspeaker response is sufficiently similar to the direct sound in the reflected sounds despite the directionality of the loudspeaker and reflection phenomena. The extracted direct sound is thus used as the single atom in the MP dictionary. Because of the shadowing effects, the strongest of the direct sounds in the four microphones is chosen as the atom. The highpass filtering that was done earlier increases the matching pursuit's ability to find the correct waveforms and peaks in the impulse response since the ringing low frequencies of the direct sound and other reflections have a smaller effect on the waveforms and there are fewer false reflection peaks.

For a single atom, or one atom and its convolutions with different windows, the matching pursuit can be implemented efficiently without having to calculate the cross-correlation on each round [91]. The algorithm applied to searching reflections from impulse responses with a single dictionary atom is described in pseudo code in Algorithm 1. The loop is repeated N number of times. There can also be a maximum number of reflections $n_{\max}$ after which the algorithm is stopped. If both parameters are set, the maximum number of iterations should be larger than the maximum number of reflections since many iterations might not produce new found reflections due to overlap with a previous one. The maximum number of reflections should be in most cases set significantly higher than the maximum number of reflections possible to model since some of the reflections found in the MP will also be discarded later in the analysis. An example of reflection times τ_j found using Algorithm 1 is shown in Figure 11.

The time constant $t_{\min}$ in Algorithm 1 is used to reject peaks in the cross-correlation too close to already found reflections and is important for the later time delay estimation. If two reflections are found too close to each other they might be part of the same reflection and if they are separate overlapping reflections, it is difficult to determine which reflection in one impulse response corresponds to which reflection in the impulse responses of the other microphones. However, discarding these peaks in the cross-correlation means that overlapping reflections are not all analyzed, only the strongest one is. Due to the imperfect omnidirectionality of the capsules, there is a risk that different reflections may be strongest in different microphones leading to TDOA analysis between two different reflections in two different microphones. The direction of arrival would in this case be totally unpredictable.

Algorithm 1 Modified Matching Pursuit

calculate the autocorrelation of the atom:
$$\Psi[k] = \sum_{l=0}^{L-k-1} x_{atom}[l]x_{atom}[l+k]$$
calculate the cross-correlation of the atom and the impulse response $x[k]$:
$$\chi[k] = \sum_{l=0}^{L-k-1} x[l]x_{atom}[l+k]$$
initialize iteration round i, the number of found reflections j:
$i \leftarrow 1$
$j \leftarrow 0$
while $i < N$ **and** $j < n_{\max}$ **do**
  find the delay τ that maximizes the cross-correlation $\chi[k]$
  $\chi[k] \leftarrow \chi[k] - \chi[\tau]\Psi[t-\tau]$
  $reject \leftarrow$ **false**
  for $k = 1$ **to** j **do**
    if $|\tau - \tau_k| < t_{\min}$ **then**
      $reject \leftarrow$ **true**
    end if
  end for
  if $reject$=**false then**
    $j \leftarrow j + 1$
    $\tau_j \leftarrow \tau$
    add $\chi[\tau_j]x_{atom}[k-\tau_j]$ to the set of reflections
  end if
  $i \leftarrow i + 1$
end while

A scaled and delayed version of the direct sound atom might not represent the reflection very accurately nor give precise enough results in time delay estimation. Therefore, it is better to use a segment of the actual impulse response corresponding to the atom. Because of overlapping reflections and the diffuse energy present in the impulse response taking simply the time segment span by the delayed atom might cause errors. However, taking the envelope of the atom, smoothing it and using this as a window to cut the desired segment out of the impulse response seems to attenuate the effects of the surrounding impulse response content and isolate the parts of the waveform essential for time delay estimation. The smoothed envelope is calculated in the same way as in the direct sound extraction. A comparison of a delayed and attenuated version of the direct sound and the windowed impulse response segment is shown in Figure 12.

Upsampling

Upsampling to doubled sampling rate was applied before the matching pursuit algorithm and it was noted that in many cases it increased the number of found reflections. This is likely due to the advantages the upsampling gives for the cross-

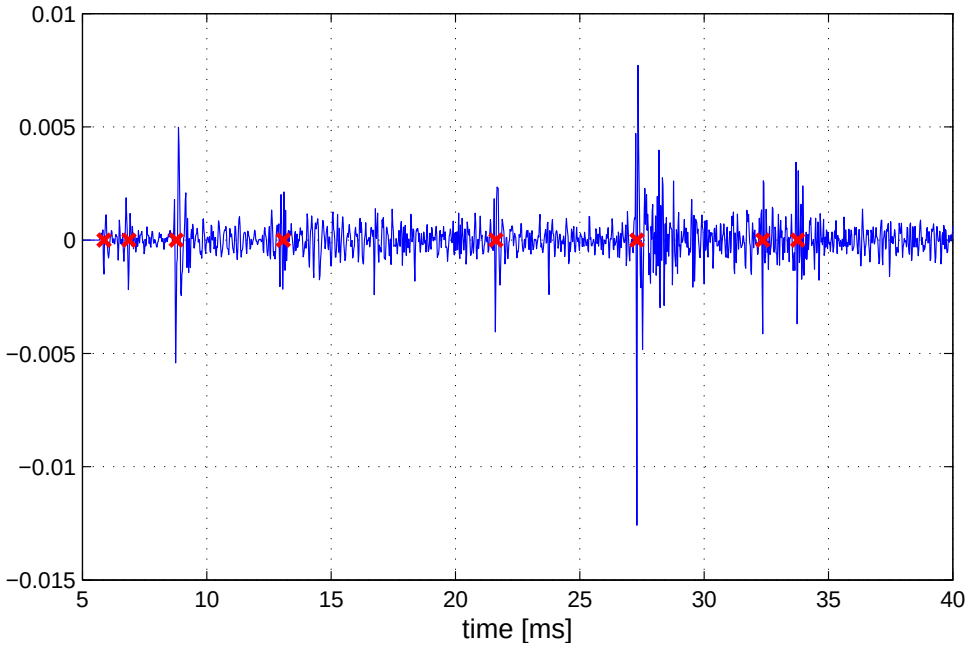

Figure 11: A section of the room impulse response with found reflections marked with red crosses.

correlation in the matching pursuit. The impulse-like reflections get slightly spread during the sampling process if their peaks happen to be between the samples. Due to the effects of the sampling their overall waveform might differ from that of the direct sound. This will cause the peaks in the cross-correlation to spread. The reflections might not then be found because the matching pursuit finds the highest peaks in the cross-correlation. Upsampling leads to smoothed waveform variations from sample to sample and reduces the limitations of the time-alignment of the direct sound and the reflections, hence making it more likely for the reflections to show up as high-amplitude peaks in the cross-correlation which will also make them more likely to be found in the matching pursuit.

6.3 Grouping the Found Reflections

The extraction of reflections using the MP algorithm is done separately for all microphones. The output of this process is a large number of impulse responses of reflections. For the time delay estimation and thereby for source localization, the representations of the reflections at each microphone must be grouped so that they can be compared. From this point on, one of the microphones will be used as the primary microphone, a time reference to which the other ones will be compared. Basically any of the microphones can be chosen to be the reference.

Matching and grouping the reflections at the microphones starts by ordering the reflections at the reference microphone by their arrival times which are estimated by the peak in their envelopes. For each found reflection at the reference microphone,

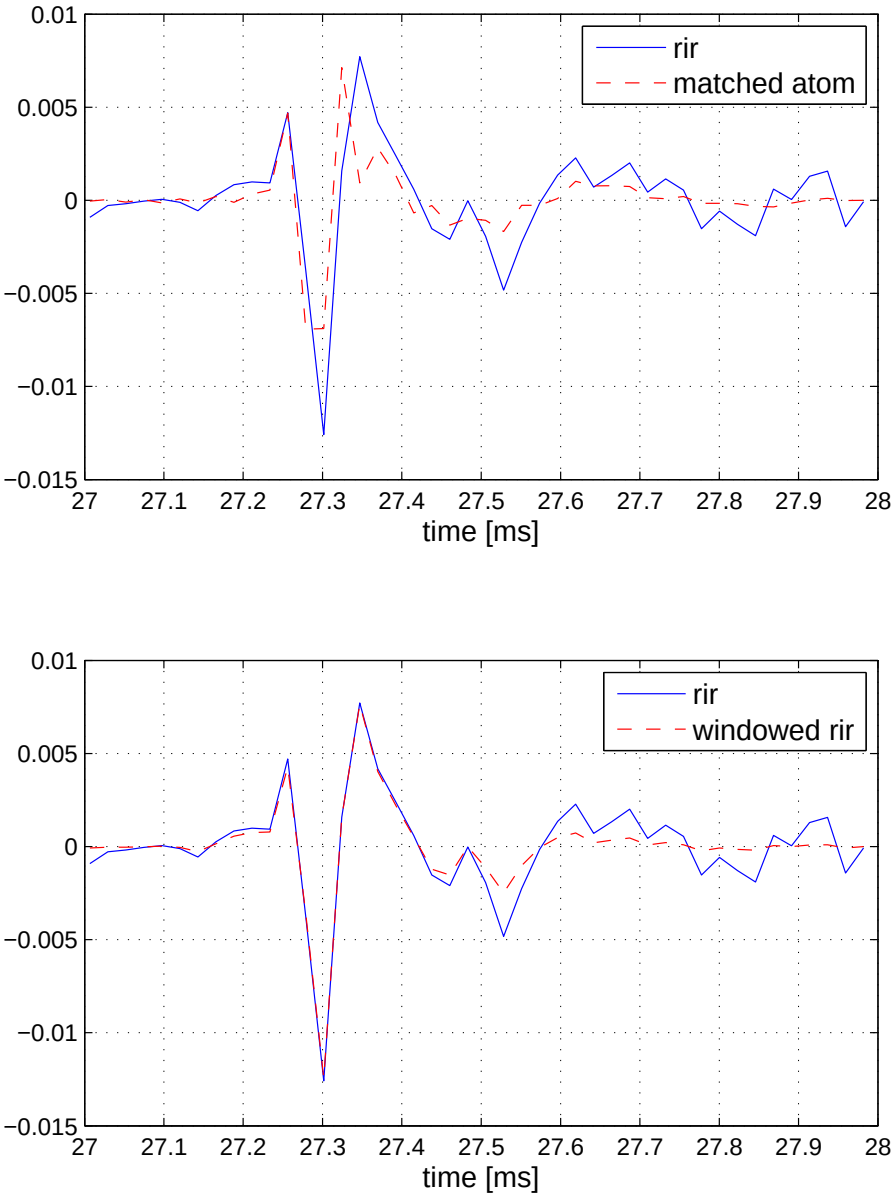

Figure 12: The representation of the reflection from the matching pursuit is a delayed and scaled version of the direct sound (upper figure). A more accurate representation is found by windowing the impulse response with the smoothed envelope of the scaled and delayed direct sound (lower figure).

the reflection closest in time at each of the other microphones is assumed to be the corresponding reflection in that microphone. Because of noise and inaccuracies in the previous processing, several constraints and error checks are necessary to make the grouping robust.

First of all, the time difference between the microphones must not be longer than the physical limits of maximum delay defined by the array geometry. The physical maximum value of the time delay for microphones spaced at distance d from each other is

$$\tau_{\max} = \frac{d}{c}. \tag{34}$$

The delays between the microphones used here are not to be confused with the delays used for source localization. These are just very rough approximates of the delays which is to say that the time span inside which they are allowed to be must be slightly larger than the physical limit of the delay.

Depending on the parameters of the matching pursuit $(N, n_{\max}, t_{\min})$ there can be very faint matches to the direct sound and the signals at some microphones might not be reliably localized. Therefore, thresholds for the strength of the extracted reflections are set. The reflections must be strong enough in all the microphones and one of them must be even stronger so that there is one to be used in the spectral analysis. The strengths of the direct sound and reflections used here are the signal energies:

$$E = \sum_{i=0}^{L-1} |x[i]|^2 \tag{35}$$

The thresholds used in the implementation were 40 dB below the direct sound in all the microphones and 30 dB in at least one of them.

6.4 Source Localization

Once the direct sound and reflections are extracted in each impulse response and grouped together, TDOA-based closed-form source localization is used to find the source and image source locations. The direct sound is treated similarly to the reflections in the localization analysis.

The TDOA estimation is done using GCC with the simple cross-correlation weighting. As explained in Chapter 5, the peak of the GCC presents the time delay. Parabolic interpolation is used for the peak and two of its surrounding samples to get a more precise peak location estimate (see Figure 13). The delays are calculated between the reference microphone and the three other microphones leading to three TDOAs for each reflection.

In order to avoid problems in the localization, TDOAs outside the physical limits of the time delays, as defined in Equation (34), are rejected. Once again due to estimation errors and the possible inaccuracy of the assumed speed of sound, the rejection threshold is set a little above the calculated physical limit τ_{max}. This time the rejection threshold time is set shorter than in the grouping phase of the analysis since these delay values are the actual ones used to localize the sources.

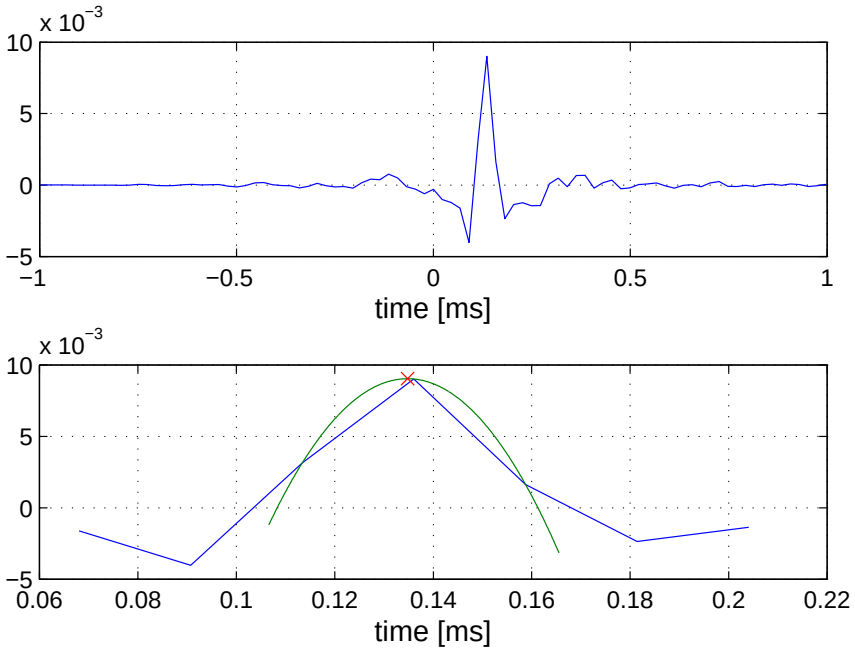

Figure 13: The time difference of arrival between two microphones estimated as the cross-correlation of the signals (upper figure). A more accurate peak location is found by parabolic interpolation around the peak (lower figure). The fitted parabola is shown in green and its peak marked with a red cross.

After the TDOAs have been estimated, Equation (32) is used to calculate the source locations. The coordinates of the sources depend on the chosen coordinate representation of the microphones with respect to the reference microphone's location which is the origin of the system and hence also on the orientation of the microphone array during the measurement. As a post-processing step after the localization, the sources can be rotated around the origin keeping their relative positions constant. This enables the front-facing head orientation in the auralization to be independent of the orientation of the microphone array during the measurements. The analyzed location of the center speaker or a stereo speaker pair can be used as a rotation reference which is set to be straight in front of the listener in the auralization.

6.5 Choosing Image Sources for Synthesis

Due to hardware limitations and typically small perceptual importance of individual high order reflections, it is necessary to choose which of the analyzed image sources are most important. Although in many cases the analysis does not find more sources than it is possible to model, this step is especially necessary for cases where the reflections are strong, dense and arrive early. In this work, the strongest reflections in terms of their energies (see Equation (35)) are given the highest priority. The

energies are averaged over the four microphones. Weaker reflections are left out of the image source model depending on the computational limitations on the number of reflections.

For future development, several criteria for rejecting or combining sources due to modeling limitations exist. If the number of modeled image sources is limited, sources close in space could be combined, for instance by replacing them with a single source that has an intermediate position and summed energy over those sources. In this case, the sources must actually be close enough in space, not just have a similar direction of arrival. In other words, it is perceptually justified to combine sources only if their arrival directions and arrival times are both close enough. In addition to combining sources, they can be left out of the modeling if they are assumed to be masked by other reflections. Once again, times and angles of arrival must be taken into account along with the energies of the reflections. Plenty of research exists about the audibility of reflections (see e.g. [93, 94, 95, 96]). The masking thresholds and just noticeable differences presented in these studies could serve as a basis for creating criteria for the rejection and combination rules.

6.6 Spectral Analysis

After determining the source locations, the image sources should be parameterized with propagation filters and gains. Propagation filters account for the directivity of the sources and material absorption as was discussed earlier in Chapter 3. The on-axis frequency responses of the sources are not aimed to be analyzed and reproduced. Instead, the user is given optimal loudspeakers in the auralization by not using any loudspeaker impulse responses for the direct sounds.

The parameterization of the propagation filters requires the comparison of the magnitude responses of the reflections and the direct sound. If this can be reliably done, the filters should in theory model all the magnitude response effects on the propagation path. In addition to the previously mentioned directivity and material absorption, air absorption, effects of diffusive reflections and diffraction should also appear in the frequency response derived from the measurements and thus these effects would also be modeled unlike in the case of hand-set parameters.

However, there are several problems related to the analysis process. First of all, typically in loudspeaker impulse responses, the low frequencies get delayed more than the high frequencies [88] and the low frequencies overlap with the first reflections. Therefore, in the extraction of the direct sound, some of the low frequency information is lost. The low frequency response of the loudspeaker is thus hard to estimate for the both direct sound and the reflections. Several techniques have been developed to recover the low frequency response [97, 98, 99, 100, 101, 102, 103, 104] but they do not seem to be applicable here since the filtering techniques [97, 101] require knowledge of the speaker and and parametric spectral analysis [98, 99] does not work well for noisy responses.

The mounting of the microphone capsules shadows some frequencies of the incoming waves and the capsules themselves start getting directional at high frequencies. Using always the microphone with the smallest angle relative to the incoming

sound, the maximum angle at which the sound has been recorded is 71 degrees. Therefore, the lowpass effects of the shadowing should not be significant considering the geometry of the mounting. Another problem is the presence of diffuse energy in the reflection responses and the overlap of several reflections. Beamforming approaches might thus be better for reflection response estimation but would require larger microphone arrays and more complex processing to be effective.

Since recovering the low frequency information failed in experiments, simplified propagation filters were designed based on the data on the available frequency range. The frequency responses of the direct signal and reflection are calculated in Bark bands (listed in Table 2). The frequency resolution limit Δf of the analysis of the truncated response is $\Delta f = 1/T$ where T is the length of the response [101]. Therefore, the parts of the magnitude responses below $f_{\text{low}} = 1/T$ are given the magnitude at the lowest Bark band entirely above the low frequency limit f_{low}. The Bark bands are used in order to smooth the frequency responses. Bark bands correspond approximately to the critical bands of the human hearing and thus the smoothing increases reliability of the analysis while maintaining all perceptually relevant information. Nowadays, ERB (equivalent rectangular bandwidth, [105]) bands are more commonly used but Bark bands were used here since they were already in use in existing parts of the system.

Since this propagation filter analysis is not necessarily very reliable, the propagation gains are analyzed separately and the average magnitudes of the filters are set to zero. Distance-dependent gains are modeled in the ISM with separate gain factors in any case which means that the gains analyzed here can be set to replace distance-dependent gains and no additional parameters are needed. Because of the separate gains, the filters are set to be optional so that they can be left unused in case the filter approximations seem unreliable based on visual inspection of figures provided by the analysis system. The gain G_i of reflection i is calculated by comparing the average gain of the reflection at the four microphones to the average gain of the direct sound at the four microphones:

$$G_i = \frac{\frac{1}{4}\sum_j \sqrt{E_{ij}^{\text{reflection}}}}{\frac{1}{4}\sum_j \sqrt{E_j^{\text{direct}}}}, \tag{36}$$

where $E_{ij}^{\text{reflection}}$ is the energy of the reflection i and E_j^{direct} is the energy of the direct sound at microphone j.

Table 2: Bark bands as listed in [106]. The last Bark band is in practice extended up to the Nyquist frequency.

Bark band	center frequency (Hz)	limits (Hz)
1	50	0-100
2	150	100-200
3	250	200-300
4	350	300-400
5	450	400-510
6	570	510-630
7	700	630-770
8	840	770-920
9	1000	920-1080
10	1170	1080-1270
11	1370	1270-1480
12	1600	1480-1720
13	1850	1720-2000
14	2150	2000-2320
15	2500	2320-2700
16	2900	2700-3150
17	3400	3150-3700
18	4000	3700-4400
19	4800	4400-5300
20	5800	5300-6400
21	7000	6400-7700
22	8500	7700-9500
23	10500	9500-12000
24	13500	12000-15500

7 Late Field Analysis in the Implemented System

7.1 Reverberation Time

The FDN used in the late field modeling can be controlled with various parameters. The key parts for matching the measured room's late field characteristics are the frequency-dependent reverberation time and the overall frequency response. The analysis of these properties from the measured impulse responses follows mostly the analysis described by Jot [107, 7].

In the developed system it is enough to have the reverberation time as experienced at the static listener position which means that we can use the impulse responses measured at the listener position. The directionality of the source is not a problem because it is a desired property also in the auralization.

The STFT of the RIR is calculated using the same parameters as in [7]: the window length is 16 ms and window overlap 75%. The STFT is smoothed in Bark bands (see Table 2) and reverberation times are calculated in each of the bands using the EDR (see Section 4). Examples of the STFT and the EDR of the same RIR are depicted in Figures 14 and 15, respectively. The recirculation filters for the FDN can be designed based on the the Bark band reverberation times [107]. The gain of the filter is derived from the length of the corresponding delay line i and the reverberation time:

$$\frac{G_i(f)}{-60dB} = \frac{l_i/f_s}{RT_{60}(f)} \tag{37}$$

$$G_i(f) = \frac{-60l_i}{f_s RT_{60}(f)} dB \tag{38}$$

where l_i and $G_i(f)$ are the length of the delay line i and the frequency-dependent gain of the cascaded filter, respectively.

For this application we are interested in the characteristics of the actual late field only since early reflections are separately simulated. Therefore, the start time of the reverberation time estimation must be carefully chosen. Using the NED value 1.0 (calculated using a 25-ms Hanning window) and the result of the regression formulas (17) and (18) as the start time both gave bad estimates for the responses available for evaluating the analysis system. These estimates might be accurate estimates for the actual diffuse field start time but they are too late to be used in the reverberation time analysis because they are relatively close to the point where the noise floor is often reached and therefore leave a very short segment of the decay curve for the reverberation time estimation. An NED value of 0.7 was settled on as it performed well with the responses available. In order to increase robustness, the time average between the times when 0.65 and 0.75 are reached was used in the implementation. An example of an estimated late reverberation starting point is depicted in Figure 16.

Jot's iterative procedure (see Chapter 4) for reverberation time analysis starts with the calculation of an initial estimate of the reverberation time. In the responses fed to the parameterization system, the actual impulse response may constitute only a very short part of the measurement, the rest being only a recording of the noise

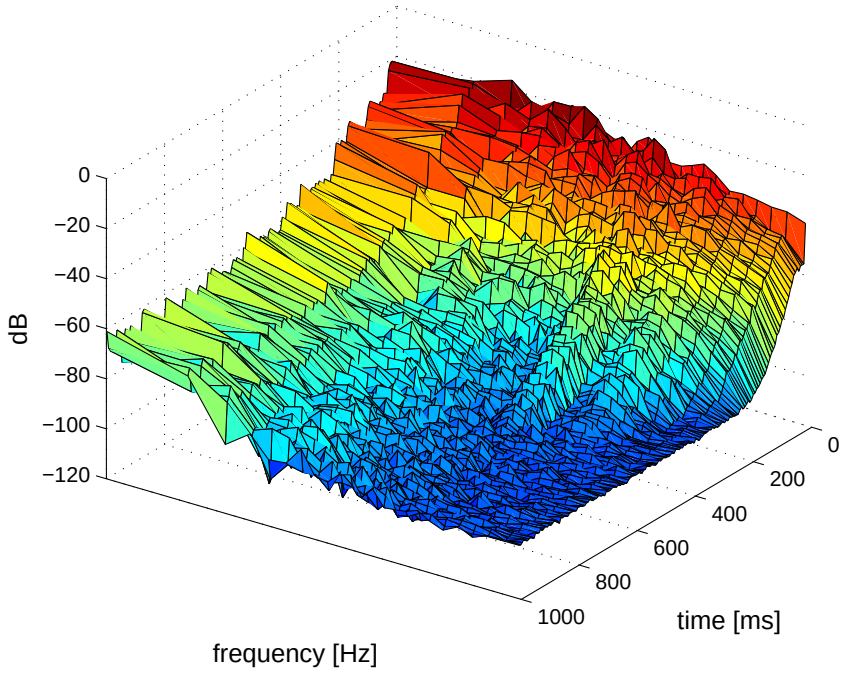

Figure 14: An example of a short-time Fourier transform of a measured room impulse response averaged in Bark bands.

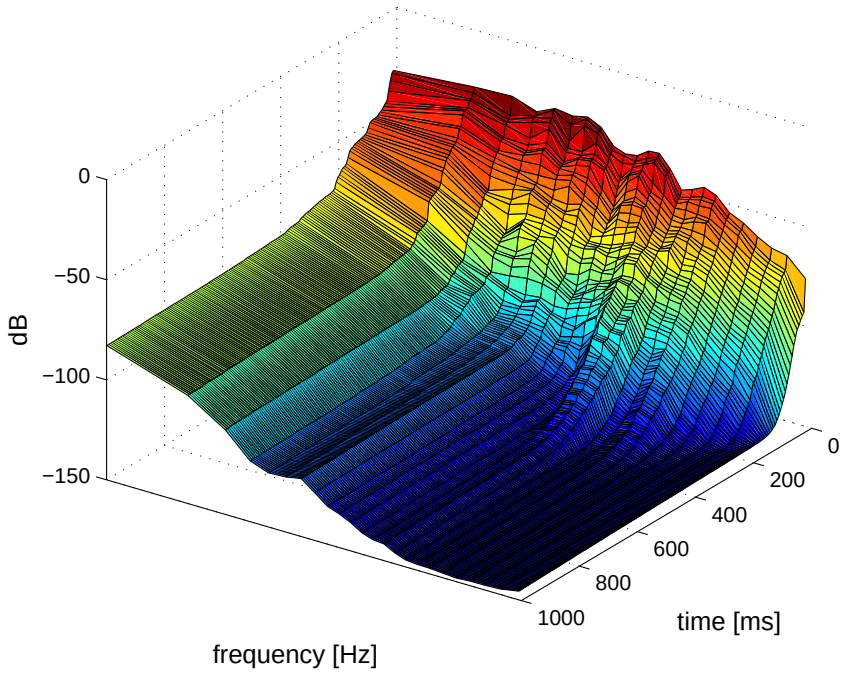

Figure 15: Energy decay relief calculated by backwards integrating in the frequency bins of the short-time Fourier transform seen in Figure 14.

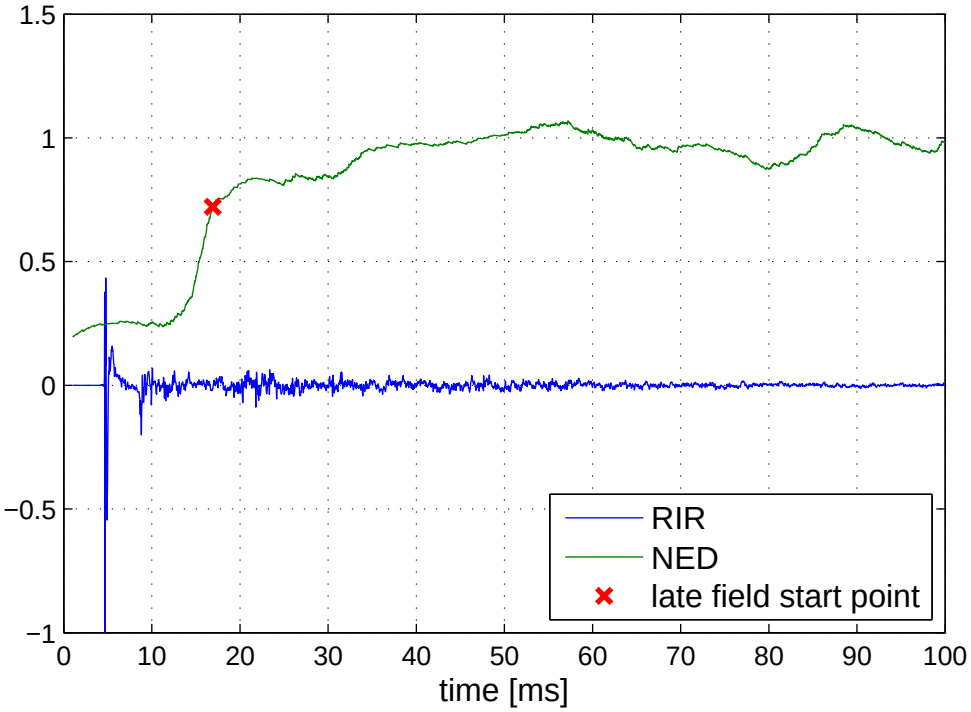

Figure 16: A room impulse response, its echo density profile calculated using normalized echo density with a 25-ms Hanning window and the reverberation analysis starting point calculated as the average of the times when normalized echo density reaches values 0.65 and 0.75.

floor. As mentioned in Chapter 6, the only assumption made about the reverberation time with respect to the length of the measurement is that the measurement is long enough so that the latter half of the measurement is only noise. The initial estimate of the reverberation time would thus be calculated from the first half of the measurement and the estimate may be very far from the real reverberation time due to the possibly long noise floor part used in the calculation. A non-iterative simplified version of Jot's procedure was applied because of this issue with the initial estimate and because the simplified version worked well enough in practice. The analysis is done in each Bark band separately. First, a linear scale line is fitted to the end part of the response where only noise is present and this line is subtracted from the entire response. The ending point for the reverberation time analysis is set approximately to the point where the decay first reaches the noise floor. This is approximated by the point where the reverse-integrated decay curve including noise reaches a level 3 dB above the estimated noise floor. The effect of approximate noise removal in the EDR and the reverberation time analysis limits are depicted in Figure 17.

The reverberation time estimates in Bark bands are calculated from each of the $N_{\text{microphones}}N_{\text{sources}}$ impulse responses separately. Averaging over these estimates gives the final Bark band reverberation times. Setting strict time limits to the

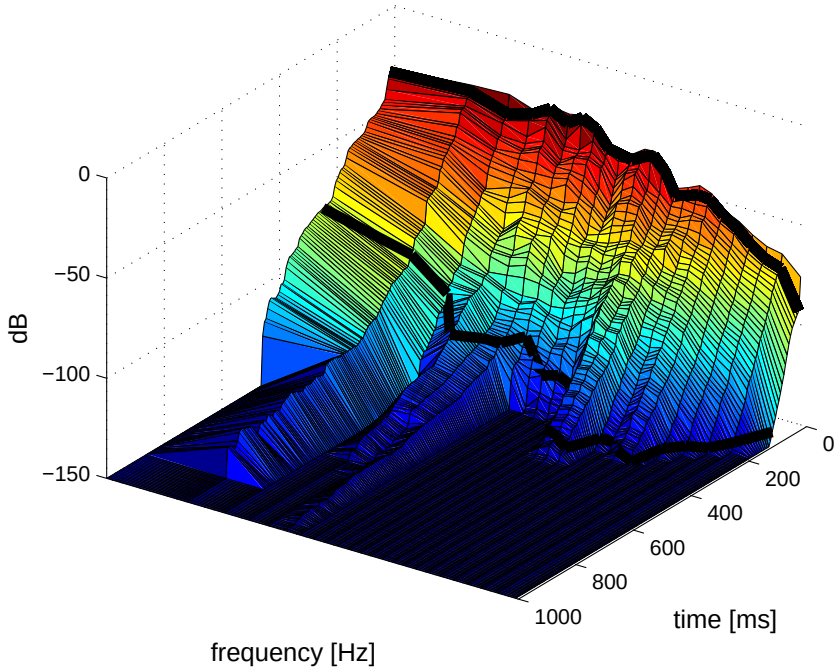

Figure 17: Energy decay relief of Figure 15 with noise approximately removed in all Bark bands separately. The reverberation times are estimated in each frequency band separately using only data within the time limits shown with the dark lines. The first limit is calculated using normalized echo density (see Figure 16) and the ending limit is the point when the original energy decay relief reaches a level 3 dB above the estimated noise floor.

analysis leads in some cases to estimation based on very little data. To avoid these unreliable estimates, we set a requirement for the dynamic range on which the line fit is done. If the dynamic range is less than 20 dB, the reverberation time estimate in the corresponding frequency band is not used. If none of the responses has enough power in a certain frequency band to create a reverberation time estimate, the estimate for that band is taken as the average of the surrounding bands.

7.2 Equalization

The recirculation filters parameterized as described above account for the frequency-dependent decay of the reverberation. The overall frequency response and level of the FDN output with respect to the direct sound has to be separately adjusted. This correction is done outside the FDN structure with a separate equalizer. The equalizer is designed by comparing the measured energy of the late reverberation and the energy of the FDN output in Bark bands. The basic idea is that the ratio of the frequency-dependent energies of the FDN output $E_{\text{output}}(f)$ and input $E_{\text{input}}(f)$ should match the ratio of the frequency-dependent energies of the late field $E_{\text{late}}(f)$

and the direct sound $E_{\text{direct}}(f)$ in the measurement after equalizer $H_{\text{EQ}}(f)$:

$$H_{EQ}(f)\frac{E_{\text{output}}(f)}{E_{\text{input}}(f)} = \frac{E_{\text{late}}(f)}{E_{\text{direct}}(f)}. \tag{39}$$

Setting the input of the FDN to an impulse with amplitude 1.0 whose energy at all frequency bands is also 1.0 leads us to the desired frequency response of the equalization filter:

$$H_{EQ}(f) = \frac{E_{\text{late}}(f)}{E_{\text{output}}(f)E_{\text{direct}}(f)}. \tag{40}$$

The energy of the direct sound is calculated as it was done in the spectral estimation of the reflections (Section 6.6): the frequency response of a real measurement loudspeaker retrieved from the strongest extracted direct sound (one per source) is averaged in Bark bands and the bands below the low frequency limit f_{low} are given the same magnitude as the lowest Bark band above f_{low}. The band-limited nature of the frequency response estimate makes the compensation inaccurate.

The energies of the late reverberation, in both measured responses and FDN output, are calculated using the STFT with the same parameters as in the reverberation time analysis. The magnitude spectrums calculated with STFT are averaged in Bark bands on the dB scale. The energies at different Bark bands are calculated by summing the linear scale magnitude values between the late field start time and the time when the noise floor is reached both of which are defined in the reverberation time analysis. The corresponding analysis for the FDN output is done for the same length of response but starting from the beginning of the FDN output because the starting point of the FDN in the auralization is set to be at the measured late field start time (based on NED value). This procedure should lead to matching the energies of the measured late field and FDN late field in the same time span when regarding the direct sound arrival as the zero time.

For the energy analysis, there are $N_{\text{microphones}}N_{\text{sources}}$ impulse responses. The 12-delay-line FDN can be excited from 12 different input channels of which N_{sources} are used in the auralization. The measurements include excitation from one source per response and thus the FDN is also excited for each input separately. The late field energies of the measurements are divided separately by the direct sound of the corresponding source leading to a set of desired frequency responses of the FDN. An average of these frequency responses is divided by the average of the responses of the FDN. The outcome is the desired response of the FDN equalizer in Bark bands. An 8th order IIR filter is designed using the Yule-Walker algorithm.

Informal listening of the auralization and comparing it to loudspeaker listening showed that the late field created by the automatic parameterization system did not quite match the late field of the room. This problem was especially strong in cases where the loudspeakers were far enough for the late field to be dominating with respect to the direct sound. According to listening and measurements, this seemed to be due to an imprecise level and frequency response of the modeled late field. The difficulty in setting the level is at least partly due to the fact that the feedback structure of the FDN does not allow fade-in. The FDN is introduced to

the auralization with a delay corresponding to the measured mixing time and decays according to the measured reverberation times. Matching the level of the FDN in the late reverberation part leaves the overall energy of the modeled reverberation too low since the diffuse field already present during the early reflections is missing. Compensating this lack of reverberation energy by raising the level of the FDN would cause the FDN to start with an energy too high compared to the early reflections. Another problem is caused by the imprecise frequency response estimates of the direct sounds making it difficult to analyze the required equalization with sufficient precision. Consequently, despite efforts to enhance the system, the problem remained and thus an additional half-automatic processing step was added until future development would lead to more exact late field equalization methods.

The additional processing step requires the impulse response measurement of the auralization system. The late field part of the impulse response is compared to the corresponding time interval in the original real world measurement. The difference of the energies of the two impulse response segments are calculated and used in a new equalizer design as an additional overall gain which reduces the gain offset between the real world target late field and the modeled late field.

8 Listening Test

For evaluating the analysis system, simulations would have been beneficial in giving information on the precision of the various parts of the analysis system. However, during the development of the algorithms and their testing it became obvious that the parameterization system had to be prepared for complex and unpredictable acoustical phenomena, such as the impulse responses and polar patterns of various loudspeaker types, diffraction from loudspeaker stands and other small objects as well as arbitrary absorption effects of wall materials. Testing the performance of the analysis system when encountering these phenomena would have required the implementation of an extremely precise room acoustic modeling system. The exact performance of the analysis could not be assessed even for the source localization using the measurement data available since the knowledge of microphone locations was generally not precise enough to accurately evaluate the source localization results. On the other hand, the acoustic source localization method and the reverberation time analysis method used in the system are known to perform generally well.

It was decided to confirm the system's performance approximately, e.g. by visual observations of the results of each analysis block, based on the measurements available during the development and to perform formal evaluation by a listening test. As the system is made to be used solely for auralization purposes, the perceptual performance of the system is, in any case, the most important performance measure which makes the listening test an important evaluation method.

8.1 Test Methodology

The ITU-BS.1534 recommendation [108] was used as a main guideline for the listening test. The recommendation suggests the use of the multiple stimuli with hidden reference and hidden anchor (MUSHRA) method. In MUSHRA, the test subject grades the similarity of various stimuli to a known reference on a continuous scale. The stimuli include the reference and a low quality anchor whose task is to set the low end of the grading scale.

The reference used for the evaluation of the auralization system was an actual loudspeaker setup in a room. The different stimuli to be compared to the loudspeaker reproduction were different variations of the headphone auralization. The subjects were asked to grade the headphone reproduction methods based on similarity to the loudspeaker reproduction in terms of the spatial impression. The biggest practical difficulty in this test setup was the need to alternate between loudspeaker and headphone listening. In order to place the reference into the stimulus set would have required the loudspeakers to be used while the headphones were on. Although open headphones (Sennheiser HD650) were used in the auralization, it was clear that having them on would have had an effect on the perceived audio quality. Even if their influence on the spatial image was negligible, the lowpass-like timbral effect on the loudspeaker sound might have lead the subjects to focus too much on other than spatial aspects of the sound. Therefore, it was decided that the hidden reference would be left out. Compared to the standard test procedure this does not allow

a reliable anchoring of the top of the grading scale where the hidden reference is typically assumed to be graded. The comparison of the stimuli should be possible and clear labeling of the grading scale should offer results on an absolute scale as well. Moreover, the loudspeaker listening was still used as the known reference to which the subjects were comparing the headphone reproduction methods and it was thus assumed that it defines the top of the grading scale for the subjects.

Since the goal was to evaluate the quality of the room parameterization, the stimuli to be compared was decided to be HRTFs without a room model as a low quality anchor and two versions of the room model: one parameterized automatically and the other having parameters set by hand. The manually parameterized model was given the positions of the loudspeakers and room boundaries, approximate octave-band absorption coefficients of the room boundaries and the frequency-dependent reverberation time as calculated with Sabine's formula (Equation 10).

In a pilot test, the automatically parameterized room was also presented without early reflections. The difference between the grades of the versions with and without the reflections was very small in the acoustically treated room used in the test, and thus the version without the reflections was left out of the test. The effect of the different parts of the room modeling system to the perceived quality of the auralization would have required a separately designed test. In a different type of room the audible effect of the reflections is likely to be stronger and perhaps a pairwise comparison would bring up the small differences better.

8.2 Binaural Synthesis in the Listening Test

The existing auralization system uses binaural synthesis through headphones to reproduce the sound processed with the room model. A set of non-individualized HRTFs is used to simulate direct sounds and early reflections to desired directions. As was mentioned in Chapter 1, one set of HRTFs does not give plausible results for all listeners but often individual HRTFs are not available.

In a binaural synthesis application like this, the transfer function from the headphones to the ear should be compensated for. These transfer functions do not only depend on the headphones but are also individual for each listener [109]. If individual measurements are not available and/or the system needs to work for multiple headphone models, a general diffuse-field equalization of the HRTFs can be used instead. In diffuse-field equalization, the HRTFs are divided by the average magnitude of HRTFs from all arrival angles [110]. This compensates for the transfer functions of headphones derived with the common diffuse-field calibration design principle where the headphone transfer functions are designed based on the average HRTFs over all arrival angles [109].

The lack of individual HRTFs and headphone- and listener-specific equalization limits the accuracy of the binaural synthesis but diffuse-field equalization of HRTFs provides an approximation of the necessary equalization and thus allows the current listening test setup to be used. In any case, this work relates to the room modeling part of the auralization system and the tuning of the other parts is out of the scope of this thesis. The listening test design was done so that it gives information on

the results of the automatic parameterization of the room model with respect to existing option of hand-tuning the parameters.

8.3 Subjects and Excerpts

A total of 12 subjects participated in the test. Two of them were female and ten male with ages between 20–50 years old. Two of the subjects had not done formal listening tests before.

Eight different audio excerpts were used in the test. They were chosen based on how representative they were of the typical content listened to on a surround sound system and how critical they were for revealing differences between the reproduction methods. The excerpts were all approximately 10 seconds long. Descriptions of the samples and the keywords used later are listed in Table 3.

Table 3: Excerpts used in the listening test.

	keyword	description
1	game	Includes sound of a battlefield (gunfire, shouting etc.) as discrete sound events in all channels with quiet music in the background.
2	speech	A female speech sample in English in the center channel.
3	jazz	A live jazz recording. Only ambience in rear channels.
4	noise	Pink noise played back from each loudspeaker (except subwoofer) one loudspeaker at a time.
5	stereo	Stereo pop music sample.
6	rock	A live recording of a rock band. Discrete events in all the speakers and band instruments emphasized in stereo channels.
7	classical	A classical music recording.
8	electronic	Electronic music with strong discrete sounds in all the speakers.

8.4 Test Setup

The dimensions of the listening room in the test were 7.3 m x 5.3 m x 3 m. It was carpeted and had large windows on three walls, light drywalls, some absorption elements and some diffusive structures on the walls and the ceiling. The listener position was slightly off the center of the room towards the front of the room which the listeners where facing. A 5.1 loudspeaker setup was set around the listener at a distance of 1.5 m with the stereo speakers set at $\pm 30°$, surround speakers at $\pm 110°$ and subwoofer on the floor by the center speaker. The loudspeakers used in the setup were Klein & Hummel O 110 studio monitors and a Klein & Hummel O 800 subwoofer. There was a display in front of the listener low enough not to disturb the path from the center speaker too much. The listener controlled the graphical

user interface (GUI) of the test using a mouse located on a small surface next to the listener. The subjects were asked to remain at the center of the loudspeaker setup but were allowed and encouraged to rotate their head as the auralization was head-tracked.

Before the test, the subjects were handed written instructions describing the structure of the test and the use of the GUI. The test itself had a training phase and the grading phase, as suggested in the ITU recommendation [108]. In the training phase, the subject could listen to all the experts with all the reproduction systems. The experts and the systems were grouped on the screen but their order was randomized. In the grading phase the excerpts were presented to the subject one at a time in random order. For each excerpt, the subject could listen to the reference and the three variations of the headphone auralization (ordered to A, B and C randomly for each excerpt) and grade the auralization schemes using the corresponding sliders. The subjects were able to sort the stimuli on the screen based on their current grades. A shorter segment of the excerpt could be selected for listening using a waveform view. A pop-up window reminded the subjects for putting on and taking off the headphones between headphone and loudspeaker listening. This added delay between the playback through loudspeakers and headphones but made it unlikely that a subject would listen to the loudspeakers through the headphones which might change the perceived spatial image as was stated above.

8.5 Results

The average scores and their 95% confidence intervals for each sound excerpt and reproduction scheme are depicted in Figures 18 and 19. The average scores and 95% confidence intervals for the reproduction schemes over all the excerpts are depicted in Figure 20. The reproduction with plain diffuse field equalized HRTFs stands out as the lowest quality system with most content types. In the classical excerpt, all the systems get quite similar scores, possibly due to the strong reverberation in the recording itself.

The differences between the automatically and manually parameterized systems are small with all the excerpts and the wide confidence intervals make the difference statistically insignificant. The scores averaged over all the excerpts (Figure 20) make the preference order of the systems more clear showing the automatically parameterized system to outperform the manually parameterized system. However, the confidence intervals overlap.

A statistical analysis based on mean values and confidence intervals formally assumes normal distribution of the data. The Kolmogorov-Smirnov and Shapiro-Wilk normal distribution tests showed, however, that the scores are not normally distributed. A non-parametric test was thus run for the average scores of the automatically and manually parameterized systems. The Wilcoxon pairwise comparison test [111] showed ($p=0.03$) that the automatically parameterized system performed better than the manually parameterized system.

Sporer et al. [112] discuss alternative analysis methods for MUSHRA listening tests. One of their suggestions involves calculating the difference scores between

two systems. If the confidence intervals of the difference scores do not include zero, the difference between the systems is significant. This method was applied to the scores of the automatically and manually parameterized systems (see Figure 21). The mean difference score for all the excerpts is clearly positive with its confidence intervals also fully above zero which shows again the automatically parameterized system to have performed better than the manually parameterized one.

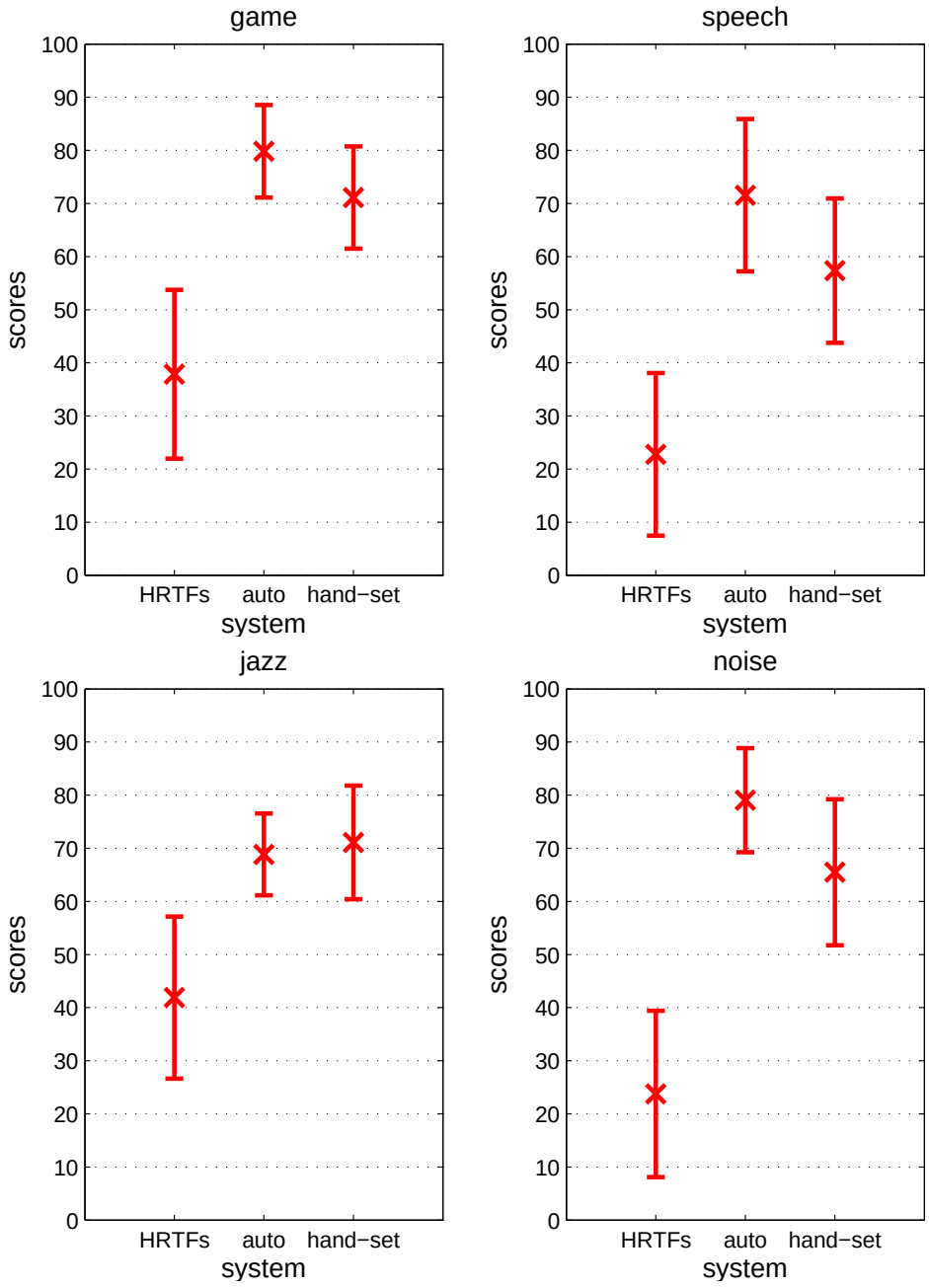

Figure 18: Average scores and 95% confidence intervals for the first four excerpts.

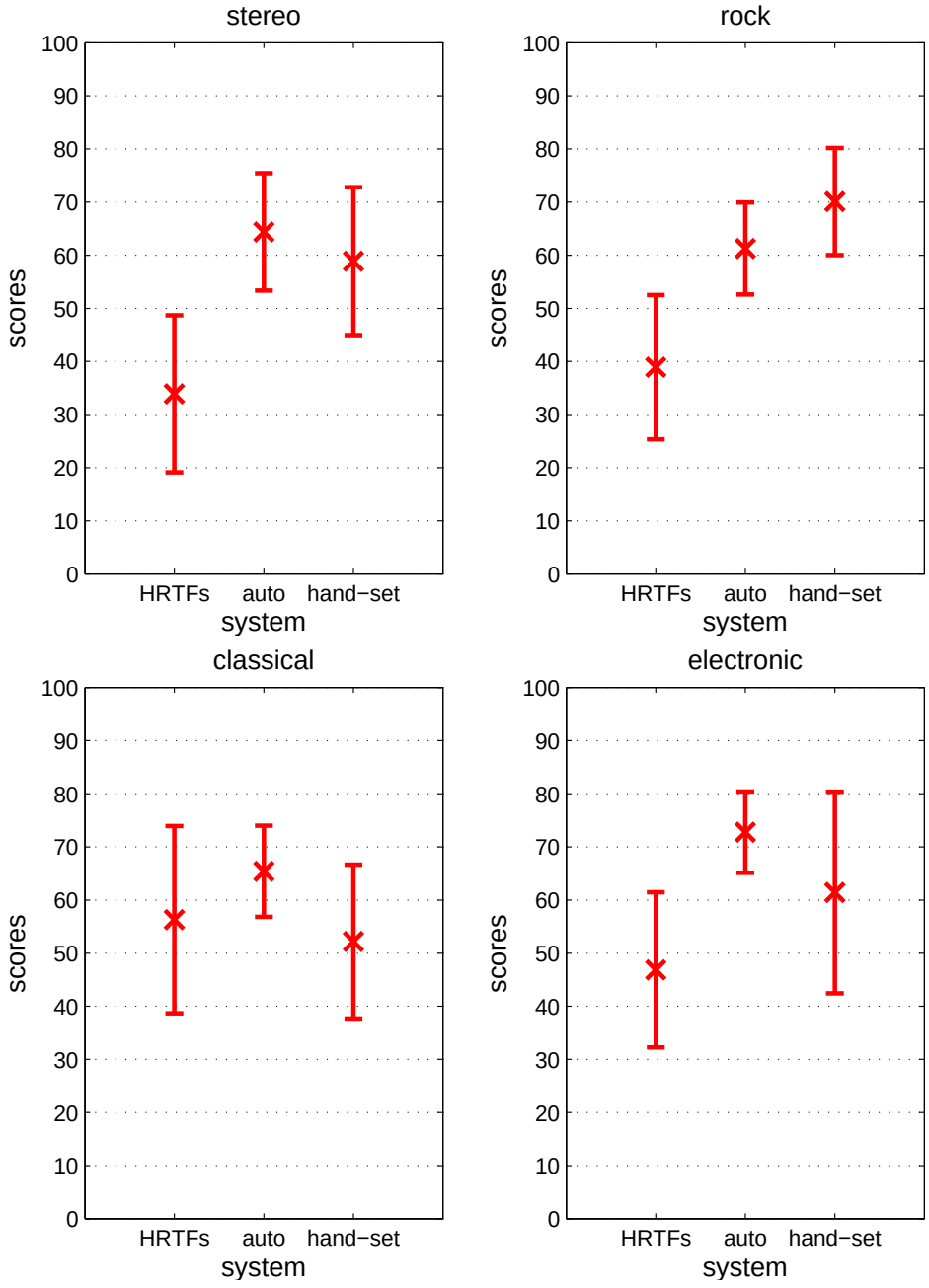

Figure 19: Average scores and 95% confidence intervals for the last four excerpts.

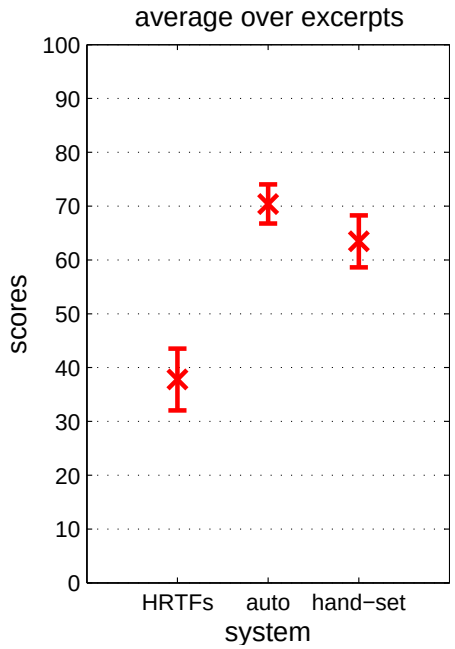

Figure 20: Average scores and 95% confidence intervals for all excerpts.

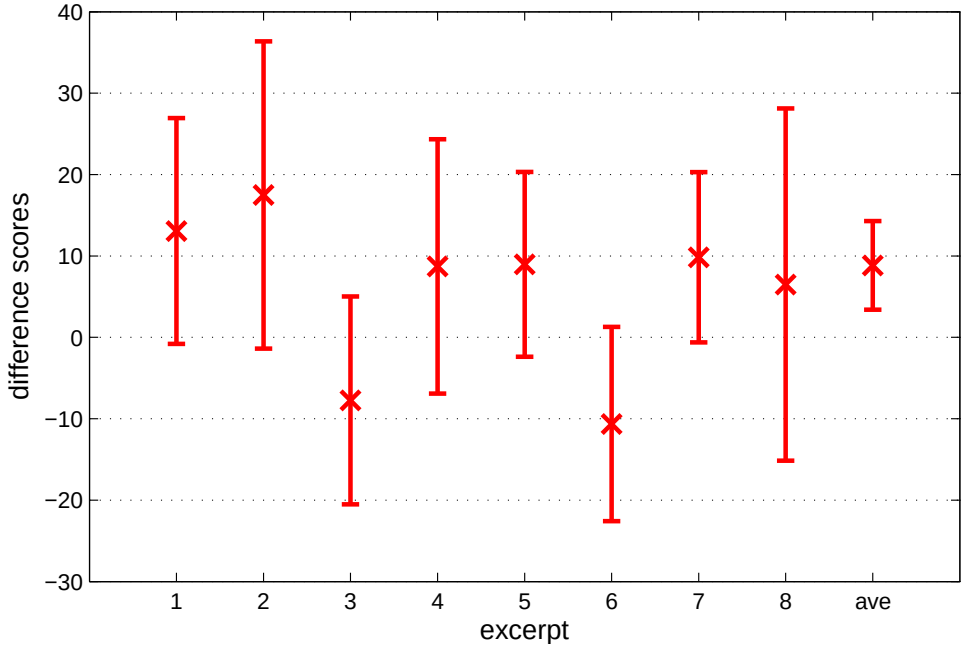

Figure 21: Difference scores for the automatically and manually parameterized auralization systems.

8.6 Discussion

Even though the listening test showed that the automatic parameterization system performed statistically significantly better than the manual parameterization and the plain HRTFs, it raised a lot of questions and possible problems. The variance in the scores was large. It is probably due to differences in perceived quality between the subjects but also due to different usage of the grading scale. The large variance can partly be explained by the number of subjects, since the ITU recommendation [108] suggests 20 subjects at minimum, but there seem to be several other problems causing this variance as well.

The reference could not be added as a high-quality anchor and so the subjects could not be told to set the best system at score 100. This is likely to have made it more difficult to set the scores for the systems. The attribute based on which the grading should be done was called "surround audio quality" in the GUI which lead some people to grade the stereo pop content very low for all the systems. In the instructions, it was said that the grading should be done based on the similarity to the reference. This should have probably been repeated on the GUI instead of using the aforementioned attribute.

One thing that made it difficult for the subjects to grade the spatial impression was the mismatches in the timbre of the sounds. Several subjects reported that some of the systems had clearly more bass than others. Yet they told they did the grading based on the spatial quality even if the timbre differences were disturbing. In any case, these comments show that the frequency response of the auralization system does not quite match the real world. This could partly be due to the fact that the impulse responses of the loudspeakers are not modeled. On the other, the loudspeakers used in the listening test have very neutral frequency responses. One important reason for the timbral issues might be the imprecise equalization of the late reverberation as discussed in Chapter 7.

Overall, it can be noticed that the perceived auralization quality is highly dependent on the content listened. The grading seemed to be clearly easier for content including discrete, dry sounds whereas reverberant sounds that are very similar in all the channels were more difficult to grade which is to say that reverberant material is less revealing for the performance of the room model.

9 Conclusions

In this thesis, a system for the parameterization of a virtual acoustic room model was developed. Relevant background theory on room acoustics, virtual acoustic modeling, room acoustic measurements and acoustic source localization was reviewed. A listening test designed to evaluate the performance of the developed system was described and results were discussed.

The developed system parameterizes the two-part room model based on room impulse responses measured with a microphone array. The parameterization system locates the acoustical sources and their reflection images. This process includes the separation of the direct sound and the detection of reflections using the matching pursuit algorithm with the separated direct sounds. Time difference of arrival -based acoustic source localization gives location estimates which, together with approximated frequency domain effects of the propagation paths, are used in the directional modeling of direct sounds and reflections in the auralization system.

The FDN used for the late reverberation in the auralization is parameterized by analyzing the frequency-dependent reverberation time, the mixing time and the overall frequency response of the measured room impulse responses. The energy decay relief analysis of the impulse responses is in a key role in the reverberation time analysis. Normalized echo density is used to find the late reverberation starting time which is used in the analysis as well as in the auralization as the input delay of the FDN. Due to the problems in the equalization of the FDN, an additional calibration step based on a measurement of the auralization system is necessary to better match the real world late reverberation.

In a formal listening test, the parameters generated with the developed system were shown to perform better than parameters set manually based on approximate geometrical information. Despite the success of the automatic parameterization, several issues remain. The equalization of the reverberation does not match the real world experience quite well enough causing, in many cases, noticeable timbral effects and reverberation too strong or too weak. The perceived quality of the auralization was also noticed to be dependent on the audio content as well as the speaker setup and the room that was modeled. The precision requirements of the parameterization system thus depend highly on the content and environments the auralization system is used for. A truly generalizable parameterization system might require different approaches for some analysis steps. Nevertheless, the developed automatic parameterization system reduces manual work, enables room acoustic modeling without expertise in the area and was shown to give better auralization results than the manual system.

In future work, a comprehensive evaluation of the individual analysis steps in different contexts is required. The experimentally derived values for various filter orders and cutoff frequencies used in the direct sound separation should be evaluated more thoroughly using data from more rooms and loudspeakers. Likewise, the values used in the matching pursuit should be tested more carefully. Alternatively, more advanced methods combining time of arrival and time difference of arrival information [87] could be applied. The equalization of the FDN might require a

different kind of approach or refinement of the actual reverberator structure used in the auralization.

This work also raised interesting questions for the general requirements of an auralization system. The importance of individual reflections and their precision, especially compared to the late field, could be investigated in different environments in order to find out what makes a plausible auralization experience. This information could be used to reduce the requirements of the modeling and focus it on the most important audible features. On the other hand, as room modeling tools get increasingly precise and their implementations cheaper, there is plenty of work to be done in order to automatically parameterize these highly refined virtual acoustic features and hence virtually recreate any existing acoustic space.

References

[1] M. Kleiner, B.-I. Dalenbäck, and P. Svensson, "Auralization – an overview," *Journal of the Audio Engineering Society*, vol. 41, no. 11, pp. 861–875, 1993.

[2] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *Journal of the Audio Engineering Society*, vol. 47, no. 9, pp. 675–705, 1999.

[3] D. Hammershøi and H. Møller, "Binaural technique – basic methods for recording, synthesis, and reproduction," in *Communication Acoustics* (J. Blauert, ed.), ch. 9, pp. 223–254, Springer, 2005.

[4] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. i: Stimulus synthesis," *Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 858–867, 1989.

[5] H. Kuttruff, *Room Acoustics*. London, United Kingdom: Spon Press, fourth ed., 2000.

[6] M. R. Schroeder and K. H. Kuttruff, "On frequency response curves in rooms. comparison of experimental, theoretical, and monte carlo results for the average frequency spacing between maxima," *Journal of the Acoustical Society of America*, vol. 34, no. 1, pp. 76–80, 1962.

[7] J.-M. Jot, L. Cerveau, and O. Warusfel, "Analysis and synthesis of room reverberation based on a statistical time-frequency model," in *Audio Engineering Society 103rd Convention*, September 1997.

[8] J.-D. Polack, *La transmission de l'énergie sonore dans la salles*. PhD thesis, Université du Maine, Le Mans, France, December 1988.

[9] P. Rubak and L. G. Johansen, "Artificial reverberation based on a pseudo-random impulse response II," in *Audio Engineering Society 106th Convention*, May 1999.

[10] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses," in *Audio Engineering Society 128th Convention*, May 2010.

[11] T. Hidaka, Y. Yamada, and T. Nakagawa, "A new definition of boundary point between early reflections and late reverberation in room impulse responses," *Journal of the Acoustical Society of America*, vol. 122, no. 1, pp. 326–332, 2007.

[12] W. C. Sabine, "Reverberation," *The American Architect and The Engineering Record*, 1900. Reprinted in Collected Papers on Acoustics. Harvard University Press, Cambridge, 1923.

[13] R. Väänänen, *Parametrization, Auralization, and Authoring of Room Acoustics for Virtual Reality Applications*. PhD thesis, Helsinki University of Technology, Espoo, Finland, May 2003.

[14] S. Siltanen, *Efficient Physics-Based Room-Acoustics Modeling and Auralization*. PhD thesis, Aalto University School of Science and Technology, Department of Media Technology, Espoo, Finland, January 2010.

[15] P. Novo, "Auditory virtual environments," in *Communication Acoustics* (J. Blauert, ed.), ch. 11, pp. 277–297, Springer, 2005.

[16] A. Pietrzyk, "Computer modeling of the sound field in small rooms," in *Audio Engineering Society 15th International Conference: Audio, Acoustics & Small Spaces*, October 1998.

[17] D. Botteldooren, "Finite-difference time-domain simulation of low-frequency room acoustic problems," *Journal of the Acoustical Society of America*, vol. 98, no. 6, pp. 3302–3308, 1995.

[18] K. Kowalczyk and M. V. Walstijn, "Modeling frequency-dependent boundaries as digital impedance filters in fdtd and k-dwm room acoustics simulations," *Journal of the Audio Engineering Society*, vol. 56, no. 7/8, pp. 569–583, 2008.

[19] L. Savioja, T. J. Rinne, and T. Takala, "Simulation of room acoustics with a 3-D finite difference mesh," in *International Computer Music Conference*, 1994.

[20] L. Savioja and V. Välimäki, "Improved discrete-time modeling of multi-dimensional wave propagation using interpolated digital waveguide mesh," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP-97.*, 1997.

[21] D. R. Begault, *3D Sound for Virtual Reality and Multimedia*, ch. 4, pp. 117–190. AP Professional, 1994.

[22] U. P. Svensson, R. I. Fred, and J. Vanderkooy, "An analytic secondary source model of edge diffraction impulse responses," *Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2331–2344, 1999.

[23] V. Pulkki, T. Lokki, and L. Savioja, "Implementation and visualization of edge diffraction with image-source method," in *Audio Engineering Society 112th Convention*, (Munich, Germany), May 2002.

[24] S. S. A. Krokstad, S. Strom, "Calculating the acoustical room response by the use of a ray tracing technique," *Journal of Sound and Vibration*, vol. 8, no. 1, pp. 118–125, 1968.

[25] B.-I. L. Dalenbäck, "Room acoustic prediction based on unified treatment of diffuse and specular reflection," *Journal of the Acoustical Society of America*, vol. 100, no. 2, pp. 899–909, 1996.

[26] A. Wareing and M. Hodgson, "Beam-tracing model for predicting sound fields in rooms with multilayer bounding surfaces," *Journal of the Acoustical Society of America*, vol. 118, no. 4, pp. 2321–2331, 2005.

[27] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[28] J. Borish, "Extension of the image model to arbitrary polyhedra," *Journal of the Acoustical Society of America*, vol. 75, no. 6, pp. 1827–1836, 1984.

[29] J. Huopaniemi, L. Savioja, and M. Karjalainen, "Modeling of reflections and air absorption in acoustical spaces – a digital filter design approach," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, October 1997.

[30] R. S. Pellegrini, *A Virtual Reference Listening Room as an Application of Auditory Virtual Environments*. PhD thesis, Ruhr-University Bochum, 2001.

[31] R. R. Torres, U. P. Svensson, and M. Kleiner, "Computation of edge diffraction for more accurate room acoustics auralization," *Journal of the Acoustical Society of America*, vol. 109, no. 2, pp. 600–610, 2001.

[32] H. Nironen, "Diffuse reflections in room acoustics modeling," Master's thesis, Helsinki University of Technology, Espoo, Finland, October 2004.

[33] T. Funkhouser, I. Carlbom, G. Elko, G. Pingali, M. Sondhi, and J. West, "A beam tracing approach to acoustic modeling for interactive virtual environments," in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques (SIGGRAPHS'98)*, pp. 21–32, 1998.

[34] S. Siltanen, T. Lokki, S. Kiminki, and L. Savioja, "The room acoustic rendering equation," *Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1624–1635, 2007.

[35] M. R. Schroeder, "Natural sounding artificial reverberation," *Journal of the Audio Engineering Society*, vol. 10, no. 3, pp. 219–223, 1962.

[36] J. A. Moorer, "About this reverberation business," *Computer Music Journal*, vol. 3, no. 2, pp. 13–28, 1979.

[37] J. O. Smith, "A new approach to digital reverberation using closed waveguide networks," in *International Computer Music Conference*, 1985.

[38] J. Dattorro, "Effect design – part 1: Reverberator and other filters," *Journal of the Audio Engineering Society*, vol. 45, no. 9, pp. 660–684, 1997.

[39] J.-M. Jot and A. Chaigne, "Digital delay networks for designing artificial reverberators," in *Audio Engineering Society 90th Convention*, February 1991.

[40] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1421–1448, 2012.

[41] J. O. Smith and D. Rocchesso, "Connections between feedback delay networks and waveguide networks for digital reverberation," in *Proceedings of the International Computer Music Conference (ICMC'94)*, 1994.

[42] J. Stautner and M. Puckette, "Designing multi-channel reverberators," *Computer Music Journal*, vol. 6, no. 1, pp. 52–65, 1982.

[43] M. A. Gerzon, "Unitary (energy-preserving) multichannel networks with feedback," *IEEE Electronics Letters*, vol. 12, no. 11, pp. 278–279, 1976.

[44] D. Rocchesso, "Maximally diffusive yet efficient feedback delay networks for artificial reverberation," *IEEE Signal Processing Letters*, vol. 4, no. 9, pp. 252–255, 1997.

[45] F. Menzer and C. Faller, "Unitary matrix design for diffuse jot reverberators," in *Audio Engineering Society 128th Convention*, (London, UK), May 2010.

[46] R. Väänänen, V. Välimäki, J. Huopaniem, and M. Karjalainen, "Efficient and parametric reverberator for room acoustics modeling," in *Proceedings of the International Computer Music Conference (ICMC '97)*, (Thessaloniki, Greece), pp. 200–203, September 1997.

[47] S. Müller and P. Massarani, "Transfer-function measurement with sweeps," *Journal of the Audio Engineering Society*, pp. 443–471, June 2001.

[48] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society 108th Convention*, (Paris, France), February 2000.

[49] "ISO 3382-1:2009. Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces." International Organization for Standardization, June 2009.

[50] "ISO 3382-2:2008. Acoustics – Measurement of room acoustic parameters – Part 1: Reverberation time in ordinary rooms." International Organization for Standardization, June 2008.

[51] M. R. Schroeder, "New method of measuring reverberation time," *Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965.

[52] J. S. Abel and P. Huang, "A simple, robust measure of reverberation echo density," in *Audio Engineering Society 121st Convention*, October 2006.

[53] P. Huang and J. S. Abel, "Aspects of reverberation echo density," in *Audio Engineering Society 123rd Convention*, October 2007.

[54] P. Huang, J. S. Abel, H. Terasawa, and J. Berger, "Reverberation echo density psychoacoustics," in *Audio Engineering Society 125th Convention*, October 2008.

[55] R. Stewart and M. Sandler, "Statistical measures of early reflections of room impulse responses," in *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-07)*, (Bordeaux, France), September 2007.

[56] G. Defrance, L. Daudet, and J.-D. Polack, "Using matching pursuit for estimating mixing time within room impulse responses," *Acta Acustica United with Acustica*, vol. 95, no. 6, pp. 1071–1081, 2009.

[57] P. Pertilä, *Acoustic Source Localization in a Room Environment and at Moderate Distances*. PhD thesis, Tampere University of Technology, Tampere, Finland, January 2009.

[58] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Springer Topics in Signal Processing, Springer, 2008.

[59] I. J. Clarke, "Efficient maximum likelihood using higher rank spectral estimation," in *Workshop on Higher-Order Spectral Analysis*, (Vail, Colorado, USA), pp. 229–234, June 1989.

[60] J. Mather, "The incremental multi-parameter algorithm," in *Twenty-Fourth Asilomar Conference on Signals, Systems and Computers*, (Pacific Grove, CA, USA), pp. 368–372, November 1990.

[61] I. J. Clarke, "Multi-signal time-frequency model fitting using an approximate maximum likelihood algorithm," in *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, pp. 269–272, October 1992.

[62] S. E. Roper, *A Room Acoustics Measurement System using Non-Invasive Microphone Arrays*. PhD thesis, University of Birmingham, December 2009.

[63] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, 1996.

[64] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[65] F. Li, R. J. Vaccaro, and D. W. Tufts, "Min-norm linear prediction for arbitrary sensor arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Glasgow, UK), May 1989.

[66] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[67] J. Chen, J. Benesty, and Y. A. Huang, "Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 1, pp. 25–36, 2005.

[68] X. Lai and H. Torp, "Interpolation methods for time-delay estimation using cross-correlation method for blood velocity measurement," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 46, no. 2, pp. 277–290, 1999.

[69] L. Zhang, "On cross correlation based discrete time delay estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, (Philadelphia, PA, USA), March 2005.

[70] P. Stoica and J. Li, "Lecture notes – source localization from range-difference measurements," *IEEE Signal Processing Magazine*, vol. 23, no. 6, pp. 63–66, 2006.

[71] Y. A. Huang and J. Benesty, eds., *Audio Signal Processing For Next-Generation Multimedia Communication Systems*. Norwell, MA, USA: Kluwer Academic Publishers, 2004.

[72] T. Korhonen, *Acoustic source localization utilizing reflective surfaces*. PhD thesis, Tampere University of Technology, Tampere, Finland, October 2010.

[73] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 12, pp. 1661–1669, 1987.

[74] J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Applied Signal Processing*, pp. 1–19, 2006.

[75] J. H. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*. PhD thesis, Brown University, Providence, RI, USA, May 2000.

[76] J. Merimaa, *Analysis, Synthesis, and Perception of Spatial Sound – Binaural Localization Modeling and Multichannel Loudspeaker Reproduction*. PhD thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, Espoo, Finland, 2006.

[77] S. Tervo, "Direction estimation based on sound intensity vectors," in *17th European Signal Processing Conference (EUSIPCO 2009)*, (Glasgow, Scotland), August 2009.

[78] B. Günel, "Room shape and size estimation using directional impulse response measurements," in *Proceedings of 3rd EAA Congress on Acoustics, Forum Acusticum 2002*, 2002.

[79] D. Aprea, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic reconstruction of the geometry of an environment through acquisition of a controlled emission," in *17th European Signal Processing Conference (EUSIPCO 2009)*, (Glasgow, Scotland), August 2009.

[80] M. Kuster, D. de Vries, E. M. Hulsebos, and A. Gisolf, "Acoustic imaging in enclosed spaces: Analysis of room geometry modifications on the impulse response," *Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2126–2137, 2004.

[81] S. Tervo and T. Korhonen, "Estimation of reflective surfaces from continuous signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Dallas, TX, USA), March 2010.

[82] J. Merimaa and V. Pulkki, "Spatial impulse response rendering I: Analysis and synthesis," *Journal of the Audio Engineering Society*, vol. 53, no. 12, pp. 1115–1127, 2005.

[83] A. O'Donovan, R. Duraiswami, and D. Zotkin, "Imaging concert hall acoustics using visual and audio cameras," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '08)*, (Las Vegas, NV, USA), March–April 2008.

[84] A. Farina, A. Amendola, A. Capra, and C. Varani, "Spatial analysis of room impulse response responses captured with a 32-capsules microphone array," in *Audio Engineering Society 130th Convention*, (London, UK), May 2011.

[85] E. V. Lancker, "Localization of reflections in auditoriums using time delay estimation," in *Audio Engineering Society 108th Convention*, (Paris, France), February 2000.

[86] S. Tervo, T. Korhonen, and T. Lokki, "Estimation of reflections from impulse responses," in *Proceedings of the International Symposium on Room Acoustics, ISRA 2010*, 2010.

[87] S. Tervo, J. Pätynen, and T. Lokki, "Acoustic reflection localization from room impulse responses," *Acta Acustica united with Acustica*, vol. 98, no. 3, pp. 418–440, 2012.

[88] R. C. Heyser, "Loudspeaker phase characteristics and time delay distortion: Part 1," *Journal of the Audio Engineering Society*, vol. 17, no. 1, pp. 30–41, 1969.

[89] W. M. Hartmann, *Signals, Sound, and Sensation.* Baltimore, MD, USA: Springer, 1997.

[90] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[91] T. Collins, "Implementation of a non-linear room impulse response estimation algorithm," in *Audio Engineering Society 116th Convention*, May 2004.

[92] M. Goodwin, "Matching pursuit with damped sinusoids," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1997.

[93] S. Bech, "Timbral aspects of reproduced sound in small rooms. I," *Journal of the Acoustical Society of America*, vol. 97, no. 3, pp. 1717–1726, 1995.

[94] S. Bech, "Timbral aspects of reproduced sound in small rooms. II," *Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3539–3549, 1996.

[95] S. Bech, "Spatial aspects of reproduced sound in small rooms," *Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 434–445, 1998.

[96] D. R. Begault, B. U. McClain, and M. R. Anderson, "Early reflection thresholds for virtual sound sources," in *Proceedings of the 2001 International Workshop on Spatial Media*, (Aizu-Wakamatsu, Japan), October 2001.

[97] L. R. Fincham, "Refinements in the impulse testing of loudspeakers," *Journal of the Audio Engineering Society*, vol. 33, no. 3, pp. 133–140, 1985.

[98] E. R. Geddes, "Maximum entropy, auto regression, pole-zero modeling... on the use of modern spectral estimation in audio testing," in *Audio Engineering Society 87th Convention*, (New York, NY, USA), October 1989.

[99] P. L. Schuck, S. Olive, E. Verreault, M. Bonneville, and S. Sally, "On the use of parametric spectrum analysis for high-resolution, low-frequency, free field loudspeaker measurements," in *Audio Engineering Society 11th International Conference*, (Portland, OR, USA), May 1992.

[100] E. Benjamin, "Extending quasi-anechoic electroacoustic measurements to low frequencies," in *Audio Engineering Society 117th Convention*, (San Francisco, USA), October 2004.

[101] J. Backman, "Low-frequency extension of gated loudspeaker measurements," in *Audio Engineering Society 124th Convention*, (Amsterdam, The Netherlands), May 2008.

[102] J. Vanderkooy and S. P. Lipshitz, "Can one perform quasi-anechoic loudspeaker measurements in normal rooms," in *Audio Engineering Society 125th Convention*, (San Francisco, CA, USA), October 2008.

[103] R. Lee, "Simple arbitrary IIRs," in *Audio Engineering Society 125th Convention*, (San Francisco, CA, USA), October 2008.

[104] S. Takane, "Some further investigations on estimation of HRIRs from impulse responses acquired in ordinary sound field," in *Audio Engineering Society 127th Convention*, (New York, NY, USA), October 2009.

[105] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1–2, pp. 103–138, 1990.

[106] H. Fastl and E. Zwicker, eds., *Psychoacoustics – Facts and Models*, ch. 6, pp. 149–173. Springer, 2007.

[107] J.-M. Jot, "An analysis/synthesis approach to real-time artificial reverberation," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1992.

[108] "Rec. ITU-R BS.1534-1. Method for the subjective assessment of intermediate quality level of coding systems." International Telecommunications Union, Radiocommunication Sector, 2003.

[109] H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen, "Design criteria for headphones," *Journal of the Audio Engineering Society*, vol. 43, no. 4, pp. 218–232, 1995.

[110] J.-M. Jot, V. Larcher, and O. Warusfel, "Digital signal processing issues in the context of binaural and transaural stereophony," in *Audio Engineering Society 98th Convention*, (Paris, France), February 1995.

[111] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[112] T. Sporer, J. Liebetrau, and S. Schneider, "Statistics of MUSHRA revisited," in *Audio Engineering Society 125th Convention*, (New York, NY, USA), October 2009.